

## **Models, Simulations, and the Reduction of Complexity**

# **Abhandlungen der Akademie der Wissenschaften in Hamburg**

---

**Band 4**



# **Models, Simulations, and the Reduction of Complexity**

---

Edited by  
Ulrich Gähde,  
Stephan Hartmann,  
and Jörn Henning Wolf

**DE GRUYTER**

Die Akademie der Wissenschaften in Hamburg ist Mitglied in der



Die elektronische Ausgabe dieser Publikation erscheint seit Dezember 2021 open access.

Finanziert aus Zuwendungen der Freien und Hansestadt Hamburg.

De Gruyter

ISBN 978-3-11-031360-4

e-ISBN 978-3-11-031368-0



This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. For details go to <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

#### **Library of Congress Cataloging-in-Publication Data**

A CIP catalog record for this book has been applied for at the Library of Congress.

#### **Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.dnb.de>.

© 2013 Walter de Gruyter GmbH, Berlin/Boston

Umschlaggestaltung: Hubert Eckl, KommunikationsDesign

Redaktion: Victoria Pöhls, Elke Senne

Typesetting: PTP-Berlin Protago-TEX-Production GmbH, Berlin

Printing: Hubert & Co. GmbH & Co. KG, Göttingen

☞ Printed on acid-free paper

Printed in Germany

[www.degruyter.com](http://www.degruyter.com)

# Ulrich Gähde, Stephan Hartmann, and Jörn Henning Wolf

## Preface

In 2006, within the Academy of Sciences and Humanities in Hamburg, the working group *Models, Simulations, and the Reduction of Complexity* was founded. In this group, scientists from various disciplines – economics, engineering science, history of science, mathematics, medicine, philosophy, physics, psychology, sociology – cooperate in order to analyze methodological and epistemological problems connected with the use of models and simulations in an interdisciplinary framework. As a first public event, the members of this group, in cooperation with Stephan Hartmann (then at Tilburg University, The Netherlands), organized an international conference on *Models, Simulations, and the Reduction of Complexity* that took place at the University of Hamburg on 18–19 March, 2010. During this conference, eight selected model building and simulation projects from different disciplines from the natural, engineering, and social sciences were presented. Each presentation was commented on by a philosopher of science specializing in problems of model construction and simulation, and trained in the respective discipline. The main task of the commentators was pointing out and analyzing methodological, discipline-specific peculiarities, as well as any interdisciplinary parallels of the modeling and simulation techniques applied. The subsequent discussions focused on different strategies used for the reduction of complexity in the various disciplines, on the relation between models and underlying theories, and on the possibility for one discipline to learn from the techniques and strategies used in others. The essays and commentaries assembled in this volume are revised and extended versions of the papers and comments presented at this conference.

It is a pleasure to thank all contributors for their excellent papers and commentaries and for undertaking the task of preparing a revised version of their contribution for this volume. Furthermore, we wish to thank the Academy of Sciences and Humanities in Hamburg for generous financial and organizational support. In particular, we would like to thank its former president, Professor Heimo Reinitzer, and Dr. Elke Senne for their interest in the project and continuous support. Finally, we are grateful to Ms Victoria Pöhls for preparing the index and the final manuscript.

# Contributors

Matthias Bartelmann, Heidelberg University, Germany  
Andreas Bartels, University of Bonn, Germany  
Gregor Betz, Karlsruhe Institute of Technology, Germany  
Ralf Engbert, University of Potsdam, Germany  
Ulrich Gähde, University of Hamburg, Germany  
Martin Golubitsky, The Ohio State University, USA  
Stephan Hartmann, LMU Munich, Germany  
Dirk Helbing, ETH Zurich, Switzerland  
Robin Findlay Hendry, Durham University, United Kingdom  
Martin Hoffmann, University of Hamburg, Germany  
Tim Christian Kietzmann, University of Osnabrück, Germany  
Reinhold Kliegl, University of Potsdam, Germany  
Peter König, University of Osnabrück, Germany  
Kai-Uwe Kühnberger, University of Osnabrück, Germany  
Valerio Lucarini, University of Hamburg, Germany  
Uskali Mäki, University of Helsinki, Finland  
Wolfgang Marquardt, RWTH Aachen University, Germany  
Aleksandra Mroczko-Wąsowicz, National Yang-Ming University Taipei, Taiwan (R.O.C.)  
Julian Reiss, Durham University, United Kingdom  
Thomas Reydon, Leibniz University Hannover, Germany  
Michela C. Tacca, Heinrich Heine University Düsseldorf, Germany  
Markus Werning, Ruhr University Bochum, Germany  
Jörn Henning Wolf, Kiel University, Germany

# Content

Ulrich Gähde, Stephan Hartmann, and Jörn Henning Wolf

**Preface — V**

**Contributors — VI**

Ulrich Gähde and Stephan Hartmann

**Introduction — 1**

Matthias Bartelmann

**Cosmology – The Largest Possible Model? — 9**

Andreas Bartels

**The Standard Model of Cosmology as a Tool for Interpretation and  
Discovery — 23**

Commentary on Matthias Bartelmann

Martin Golubitsky

**Patterns in Physical and Biological Systems — 29**

Thomas A. C. Reydon

**Symmetry and the Explanation of Organismal Form — 43**

Commentary on Martin Golubitsky

Dirk Helbing

**Pluralistic Modeling of Complex Systems — 53**

Stephan Hartmann

**The Methodological Challenges of Complex Systems — 81**

Commentary on Dirk Helbing

Uskali Mäki

**Contested Modeling: The Case of Economics — 87**

Julian Reiss

**Models, Representation, and Economic Practice — 107**

Commentary on Uskali Mäki

Peter König, Kai-Uwe Kühnberger, and Tim C. Kietzmann

**A Unifying Approach to High- and Low-Level Cognition — 117**

Markus Werning, Michela C. Tacca, and Aleksandra Mroczko-Wąsowicz

**High- vs Low-Level Cognition and the Neuro-Emulative Theory of Mental Representation — 141**

Commentary on Peter König, Kai-Uwe Kühnberger, and Tim C. Kietzmann

Reinhold Kliegl and Ralf Engbert

**Evaluating a Computational Model of Eye-Movement Control in Reading — 153**

Martin Hoffmann

**Considering Criteria for Model Modification and Theory Change in Psychology — 179**

Commentary on Reinhold Kliegl and Ralf Engbert

Wolfgang Marquardt

**Identification of Kinetic Models by Incremental Refinement — 187**

Robin Findlay Hendry

**Kinetics, Models, and Mechanism — 221**

Commentary on Wolfgang Marquardt

Valerio Lucarini

**Modeling Complexity: The Case of Climate Science — 229**

Gregor Betz

**Chaos, Plurality, and Model Metrics in Climate Science — 255**

Commentary on Valerio Lucarini

**Subject Index — 265**

**Author Index — 269**

Ulrich Gähde and Stephan Hartmann

## Introduction

Modern science is, to a large extent, a model-building activity. In the natural and engineering sciences as well as in the social sciences, models are constructed, tested and revised, they are compared with other models, applied, interpreted and sometimes rejected or replaced by a better model. Some models help scientists to systematize huge amounts of data, coming from experiments or generated through computer simulation, and to extract information out of them. Other models are developed with the aim to explain a puzzling scientific phenomenon – a task that typically requires a number of clever idealizing assumptions and, more and more, the use of computer simulations. By now it is uncontroversial that scientific models are indispensable for solving scientific problems. While some philosophers (such as Ronald Giere (1999) and Bas van Fraassen (1990)) think that science can do without laws, it seems utterly impossible for science to do without models.

The extraordinary importance of models in science has not gone unnoticed by philosophers of science. Starting in the 1960s, scholars such as Peter Achinstein (1968) and Mary Hesse (1963) focused on simple models, such as the billiard ball model of a gas, to illustrate various philosophical claims about, for example, the role of metaphors and analogies in science. Others, most notably Patrick Suppes (1969), explored the connections between scientific models and mathematical (model-theoretical) models and stressed the role of models in the analysis of data (“models of data”). Later, beginning in the 1980s and initiated by seminal contributions by Nancy Cartwright (1983), Ronald Giere (1988), Ian Hacking (1983) and Bas van Fraassen (1980), increasingly complicated scientific models, from physics as well as from the special sciences, gained center stage, and new questions, for example about the relation between theories and models, came to the fore. This debate led to a rethinking of many traditional topics in the philosophy of science, including the nature of confirmation, explanation, and the structure of scientific theories, as well as the role of approximations, idealizations and intertheoretic relations. For a detailed overview of these debates, we refer the reader to the survey article by Frigg and Hartmann (2012). Bailer Jones (2009) gives a book-length discussion of models in science, including an intriguing account of the history of the philosophy of scientific models.

In order to narrow down this tremendously broad and rich field of study, we decided to focus on the modeling of complex systems. All natural and social sciences are concerned with such systems, and it is here where one of the great advantages of model-building becomes especially vivid: Modeling helps scien-

tists to make complex objects or systems comprehensible. With the help of a model, and by studying its features, scientists learn about the object or system that the model represents (van Fraassen (2008)). To model an object or system means to *reduce its complexity* and to provide a simplified description of it. This requires the identification of relevant features of the object or system under investigation that suffice, or so it is hoped, to serve a certain purpose (e.g. confirmation, explanation, prediction or understanding). This volume illustrates how this works by focusing on examples from real science, especially from bioinformatics, climate science, mathematics, neuroscience, physics, psychology, and the social sciences. We will see that the resulting equations are typically too complicated to be solved analytically, and so computer simulations are required to proceed, which stresses the pragmatic constraints on scientific models and simulations (Humphreys 2004). For further philosophical discussions of the role of computer simulations in science and the methodological problems that they raise, we refer the reader to Hartmann (1996) and Winsberg (2010).

While these questions may appear to be of exclusively philosophical importance, the practice of model-building also raises many specific methodological problems that worry scientists. Many of these problems are so specific that they are exclusively dealt with in the respective scientific community. Other questions are somewhat more general and call for a philosophical analysis; these are the ones we want to address in this volume. To do so, this volume assembles eight articles by leading scientists, each of which is commented on by a philosopher of science. At the conference, a general discussion followed. This speaker-commentator-discussion scheme led to a lively debate, and we hope that the essays assembled in this volume reflect this exchange of ideas. At this point we want to outline three major areas of discussion and interaction between scientists and philosophers of science.

First, we are interested in *descriptive questions* regarding how scientists in the various disciplines proceed when they model complex systems. Which modeling strategies do they apply? Are these strategies subject-specific, or are there more universal strategies that are useful in several disciplines? Can one scientific discipline learn from the techniques and strategies used in another?

Second, we are interested in *normative questions* of model assessment. There are several factors that play a role here and that scientists value. Scientists want, for example, that a model accounts for the available data. At the same time, they want it to be consistent with relevant theories, and they want the model to provide understanding. Note that these goals may be in conflict with each other: Models that provide understanding often do not get the data right, and, conversely, models that get the data right do not provide understanding. This raises the question of how the various goals should be weighted. And: Are these



weights subject specific, or can one say something more general here? Can the different goals of scientific modeling be reduced to one goal, say truth? These are some of the normative questions that philosophers of science address. Other, more specific normative questions, include issues regarding the empirical testing of models and the question which normative conclusions, e.g. in the social sciences, should be drawn from (typically) highly idealized models.

Third, we are interested in *epistemological and metaphysical implications* of the practice of model-building. What picture of science and the world makes best sense of this practice? Should we conclude, inspired by the apparent patchwork of theories and models, that also the world is a patchwork, i.e. shall we follow Nancy Cartwright (1999), who famously argued that the world is dappled? Or is there hope that, one day, all the bits and pieces that scientists collected will fit into a neat coherent picture of the world? And: How are theories and models related anyway? Is there a hierarchy of theoretical approaches, or do all approaches operate at the same level of fundamentality? See also Hartmann et al. (2008) and Morgan & Morrison (1999) for collections of papers on these topics.

These are only some of the questions that are addressed in this volume. Let us now shortly outline the individual chapters.

The essay by *Matthias Bartelmann* (University of Heidelberg) explains that there are at least three reasons why any attempt to construct cosmological models seems to be a bold enterprise: Firstly, we cannot do experiments with the universe as a whole. Secondly, we are part of this universe. Thirdly, we always only see a small – although growing – section of it. In spite of these challenges, the standard model of cosmology is a remarkably successful example for how the complexity of the real world can be reduced. The starting point for the construction of this model is Einstein's General Theory of Relativity. By adding two simple symmetry conditions – namely the requirements that arbitrary spatial rotations and translations should leave the observable universe unchanged for any observer – the class of so-called Friedman models is obtained. These models are characterized by a small number of parameters. It can be shown that it is possible to find a single set of parameters such that from these models a consistent picture of the actual state of the universe and its evolution can be drawn. This picture is in accordance with virtually all cosmological observations, amongst them the expansion of the universe, the cosmic microwave background and its temperature fluctuations. For this picture, however, a price is to be paid: the existence of dark matter and dark energy has to be accepted.

In his commentary on Bartelmann's paper, *Andreas Bartels* (University of Bonn) uses the standard model of cosmology as a striking example to illustrate three main tasks that models can fulfill: Firstly, they provide a net of theoretical relations that can be used to interpret data. Secondly, these data thus integrated

in the model may fulfill an evaluative function by confirming or disconfirming theoretical relations of the model. Thirdly, there is also an explorative function of models that is responsible for the research dynamics of cosmology. Models do not only describe reality, they are also instruments for exploring reality.

What happens when a phenomenon is to be investigated for which no detailed and precise mathematical model can be derived or for which a model is available but too complicated to be analyzed? These are the questions addressed in the contribution by *Martin Golubitsky* (The Ohio State University). The author shows that in such cases the existence of symmetry can nevertheless help understanding certain patterns of the system in question and enable new predictions and explanations. He illustrates this point by discussing three examples: In his first example, he discusses symmetry and symmetry breaking with respect to patterns in burner flames. The second example refers to the symmetry description of locomotor central pattern generators. Golubitsky argues that this description allows several new predictions to be made. In particular, it makes possible to predict the existence of an unexpected but natural gait shown by mammals of different species: the jump. The third example refers to the experimentally determined symmetry of the primary visual cortex. Golubitsky outlines how, through symmetry breaking arguments, an unexpected correlation between this symmetry and a variety of geometric visual hallucinations can be predicted. He thereby refers to hallucinations experienced by test persons who have taken certain drugs.

In his commentary on Golubitsky's paper, *Thomas Reydon* (Leibniz University Hannover) addresses the epistemic virtues of general mathematical models. More specifically, he asks how symmetries (as well as broken symmetries) help scientists to understand patterns exhibited by various physical and biological systems. He argues that the role of these models is more heuristic in nature: they only provide "how possibly" explanations that – when applied to biological systems – have to be supplemented by "how and why actually" explanations of functional, developmental and evolutionary biology.

In his essay, *Dirk Helbing* (ETH Zurich) is concerned with the modeling of complex systems, especially those complex systems that we find in the social sciences. This endeavor raises a number of challenging methodological questions that Helbing addresses on the basis of an analysis of a number of case studies from his own research. Helbing is especially interested in the epistemological status of multiple models for the same phenomenon. How are these different models related? Are they all true, or is none of them true? And: Is there one true model that we will develop at some point, or do we have to be content with a plurality of models? So far, science certainly confronts us with a plurality of models, and the question arises, for example, whether averaging the predictions of all such models leads to a better prediction. Helbing concludes that a paradigm shift

towards a pluralistic or possibilistic modeling approach, i.e. an approach that integrates multiple world views, is overdue and argues that it can be useful to combine many different modeling approaches to obtain a good picture of reality, even though they may be inconsistent.

*Stephan Hartmann* (LMU Munich) discusses Helbing's insights and ideas from the point of view of contemporary philosophy of science. More specifically, Hartmann distinguishes between different kinds of pluralism and elaborates on the question under which conditions the availability of multiple models is advantageous from an epistemological point of view.

In his essay, *Uskali Mäki* (University of Helsinki) focuses on the role of models in economics and the methodological questions that the practice of modeling raises. Mäki starts off by observing that models are a central tool in economics and any policy recommendation an economist gives is based on a model. At the same time, models are highly idealized and abstract from many features of the system under consideration. This prompts the question how economic models relate to the world. Or, to put it more philosophically, how does an economic model represent its target system? To address these questions, Mäki presents a detailed theory of how economic models represent an economic system and shows how this theory fits into a realist philosophy of economics that he has been defending and elaborating for many years. Furthermore, he extends his framework by distinguishing three broad ways in which modeling can be, and actually is, contested in the controversial discipline of economics. These correspond to three kinds of possible failures of modeling.

In his commentary, *Julian Reiss* (Durham University) provides a detailed criticism of Mäki's account and makes a number of suggestions for how to fix it. Reiss is especially concerned with the application of economic models to practical and policy-related problems, which raises additional methodological problems.

In their essay, *Peter König*, *Kai-Uwe Kühnberger* and *Tim Kietzmann* (University of Osnabrück) consider models of the function of the mind. This is an especially complicated task as the human mind is probably the most complex system on earth. Their ambitious goal is to present a unified model of low- and high-level cognitive systems. Such a unified model seems reasonable as high- and low-level cognitive systems implement similar structures despite their functional differences.

In their commentary, *Markus Werning* (Ruhr University Bochum), *Michela C. Tacca* (University of Düsseldorf) and *Alexandra Mrocko-Wąsowicz* (National Yang-Ming University Taipei) provide a detailed criticism of the specific model that König and collaborators suggested and discuss alternative accounts. To do so, they focus on the visual domain and draw on the theory of neuroframes.

The essay by *Reinhold Kliegl* and *Ralf Engbert* (University of Potsdam) presents an example of a model, located at the interface between experimental psychology, cognitive neuroscience, and computational neuroscience: a model for eye-movement control in reading. At a very basic level, reading can be described as an alternation between quick eye movements (saccades) and periods of relative rest (fixations). In cognitive modeling of this process, two prototypical approaches are distinguished. The serial processing approach assumes that attention moves from one word to the next, contingent on access of the meaning of a word. Also saccade programs are contingent on the completion of some lexical subprocess. In contrast, the parallel processing approach assumes that lexical and oculomotor processes are only loosely coupled and that sometimes more than one word can be processed simultaneously. The SWIFT model is an implementation of the second approach. Following up previous proposals, the authors demonstrate that the sole reliance of a criterion of goodness of fit is not sufficient for a differentiated evaluation and ranking of competing models. They illustrate the application of three additional criteria – model strictness, reliability of data, and unexpected model predictions – for the evaluation of the SWIFT model.

In his commentary, *Martin Hoffmann* (University of Hamburg) focuses on the relation between specialized models and more general and comprehensive empirical theories. He argues that Kliegl and Engbert's SWIFT model provides an example of a model that was developed largely independently of any more general psychological theory. By referring to the unexpected model predictions criterion, which Kliegl and Engbert apply to evaluate the SWIFT model, and relating this criterion to Lakatos' methodology of research programmes, Hoffmann analyzes differences in the evaluation of models and theories. He finally proposes two necessary conditions that must be fulfilled in order to turn models into useful instruments for the development of a more general underlying theory.

Within the realm of the natural sciences, models and simulations are primarily used in order to better understand and quantitatively explain natural phenomena. By contrast, modeling in the engineering sciences focuses on designing, operating and controlling artificial systems and processes. In his contribution, *Wolfgang Marquardt* (RWTH Aachen) describes the outlines of a general methodology for modeling complex kinetic phenomena that govern the behavior of chemical and biological process systems. Marquardt explains how modeling and simulation techniques, as well as techniques for model identification and discrimination are combined with high resolution-measurements in the methodology in a highly elaborate way. In an iterative process, models and underlying (real or simulated) experiments are used for mutual refinement. This methodology enables the successive design of more detailed models that are tested by an expanding data basis. Marquardt concludes by illustrating this methodology by

three typical chemical engineering problems: reaction kinetics, multi-component diffusion in liquids, and energy transport in falling film flow.

*Robin Findlay Hendry* (University of Durham) starts his commentary on Marquardt's paper by comparing the role of models in chemical engineering with the role models play in 'pure' chemistry. He then focuses on kinetic models of chemical reactions and discusses the mutual refinement of kinetic models and the corresponding experimental set-ups. He argues that the identification of the mechanisms that lead from the reactants to the products can be seen as a case of eliminative induction. Theoretical models of molecular structure and reaction mechanisms provide the starting point for this process insofar as they delimit the set of mechanisms that have to be considered.

In the last essay collected in this volume, *Valerio Lucarini* (University of Hamburg) concerns the role of models and simulations in climate science. He explains why the use of these tools faces specific problems in this field of research. One problem is related to our imperfect knowledge of the initial conditions, another to the imperfect representation of the processes of the system. Together, both deficits severely limit the possibility of providing realistic simulations and predictions. Considerable difficulties are caused by the fact that climate science does not have laboratories where models and simulations could be tested against experiments. Finally, serious methodological problems are generated by the fact that the relevant processes of climate change occur on a large variety of spatial and temporal scales. Lucarini argues that the macroscopic theory of non-equilibrium thermodynamics provides a relevant framework for improving our understanding of climate change and our ability to model it. At the same time, he rejects the expectation that there will be fundamental progress in climate science in the next few decades simply because computers become more and more powerful.

In his commentary, *Gregor Betz* (Karlsruhe Institute of Technology) distinguishes between those aspects of climate change that are known independently of any global climate model, and those that cannot be estimated without these models. He argues against the assumption that the chaotic nature of weather automatically – without any further empirical evidence – implies that the climate system is also chaotic. Given the plurality of global climate models, he then turns to the question whether and how one can empirically test, compare, and rank these rival models. To illustrate his points, Betz focuses on Lucarini's proposal of a process-oriented metrics for model evaluation, which puts special emphasis on a better understanding of the key causal processes in climate systems.

In closing, we hope that this volume will encourage the reader to reflect upon the fascinating role of models in science and that it will stimulate further discussions between scientists and philosophers of science.

## References

- Achinstein, Peter (1968). *Concepts of Science. A Philosophical Analysis*. Baltimore: Johns Hopkins Press.
- Bailer-Jones, Daniela (2009). *Scientific Models in Philosophy of Science*. Pittsburgh: University of Pittsburgh Press.
- Cartwright, Nancy (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.
- Cartwright, Nancy (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.
- Fraassen, Bas van (1980). *The Scientific Image*. Oxford: Oxford University Press.
- Fraassen, Bas van (1990). *Laws and Symmetry*. Oxford: Oxford University Press.
- Fraassen, Bas van (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.
- Frigg, Roman & Hartmann, Stephan (2012). Models in Science. In: Zalta, E. (ed.). *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*.
- Giere, Ronald (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.
- Giere, Ronald (1999). *Science without Laws*. Chicago: University of Chicago Press.
- Hacking, Ian (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.
- Hartmann, Stephan (1996). The World as a Process: Simulations in the Natural and Social Sciences. In: Hegselmann, R., et al. (eds.). *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Theory and Decision Library*. Dordrecht: Kluwer. 77–100.
- Hartmann, Stephan, Hoefer, Carl, & Bovens, Luc (eds.) (2008). *Nancy Cartwright's Philosophy of Science*. London: Routledge.
- Hesse, Mary (1963). *Models and Analogies in Science*. London: Sheed and Ward.
- Humphreys, Paul (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.
- Morgan, Mary & Morrison, Margaret (1999). *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.
- Suppes, Patrick (1969). *Studies in the Methodology and Foundations of Science. Selected Papers from 1951 to 1969*. Dordrecht: Reidel.
- Winsberg, Eric (2010). *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

**Prof. Dr. Ulrich Gähde**  
 University of Hamburg  
 Department of Philosophy  
 Von-Melle-Park 6  
 20146 Hamburg  
 Germany  
 ulrich.gaehde@uni-hamburg.de

**Prof. Dr. Stephan Hartmann**  
 LMU Munich  
 Munich Center for Mathematical  
 Philosophy  
 Geschwister-Scholl-Platz 1  
 805539 München  
 Germany  
 s.hartmann@lmu.de

Matthias Bartelmann

# **Cosmology – The Largest Possible Model?**

## **1 Laws of Nature and the Foundations of Cosmology**

Upon which foundation should one build a model for the Universe as a whole? The idea that such a model should exist seems bold itself. Can we really believe that we might be able to construct a physical model for a unique object that we cannot experiment with, that we are part of and of which we can only see a very small section? The goal of this article is to explain that this does indeed seem possible, that mathematical simplicity is used as a guiding principle in this construction, and that the resulting world model is remarkably consistent with a wealth of observations.

Let us begin with a detour through the foundation of laws of nature in physics. It is important to realise that laws of nature do not describe nature herself, but human concepts of nature. Otherwise it would not be possible to replace established laws by other, more general ones, as it has happened several times in the history of physics. Theories in physics are based on axioms chosen by physicists, and these axioms can be altered.

Newton's axioms underlie classical mechanics. They distinguish four entities; bodies, forces, space and time, and formulate how bodies move in time through space under the influence of forces. Field theory, initiated by Faraday, attaches forces to space and gives force fields their own dynamics. Special Relativity realises that space and time have no independent existence and thereby connects forces, space and time. General Relativity explains how the presence of bodies and energy affects the structure of space-time. Thus, in a general-relativistic field theory, the four initially separate entities of Newtonian physics are all linked together.

The dynamics of physical entities, i.e. their change in time, is described by differential equations. They themselves are not postulated, but derived from a more general concept overarching physics, namely that of an extremal principle. The best-known example is perhaps Fermat's principle, which states that light rays connect the source and the receiver in such a way that the light travel time is extremal along them. The extremal principle underlying essentially all of the established theories in physics is the so-called principle of least action, or Hamilton's principle. The action itself is an abstract quantity that can be constructed under very general rules. It must be independent of any observer's state of motion,



and it is typically chosen to be invariant under certain symmetry operations. Both criteria are expressed by the mathematical concept of symmetry groups.

Symmetry groups and extremal principles currently form the deepest foundation of physical laws. Which symmetry operations a physical theory should obey exactly is largely the physicists' choice. Ultimately, however, no theory is acceptable that is in demonstrable conflict with experiments.

We know four fundamental interactions. The strong force keeps the fundamental building blocks of matter bound, the so-called hadrons. The weak force is responsible for certain conversions of particles into others, in particular through the so-called beta decay. Both act only on subatomic distances. Electromagnetism keeps atoms and molecules bound and is responsible for all interactions between charges and between matter and light. Gravity is by a large margin the weakest of the four interactions. Both electromagnetism and gravity have unlimited range and thus determine physical interactions in the macroscopic world. Since the sources of electromagnetism are positive and negative charges, the effects of one type of charge can be shielded or compensated by the other type. Effectively, therefore, the range of the electromagnetic interaction is typically also limited. Shielding is impossible with gravity, as it knows only one type of charge, i.e. the mass. Electromagnetism as well as the strong and the weak interactions are described by a unified quantum field theory called the Standard Model of Particle Physics. Gravity has so far withstood all attempts to cast it into the form of a quantum field theory as well.

If we now return to cosmology, we realise that any cosmological model must essentially be derived from a theory of gravity as the only long-range force that cannot be shielded. The most advanced theory of gravity is Albert Einstein's theory of General Relativity. We must thus begin with General Relativity in our construction of modern physical world models.

General Relativity can be seen as a prototypical example for a physical theory constructed as outlined above. It is based on the fundamental concept that the geometry of space-time, characterised by its metric, is a dynamical field determined by the presence of matter and energy. The dynamical equations of General Relativity, called Einstein's field equations, follow from the principle of least action, with an action that combines the geometry of space-time with the presence of matter in what seems to be the most straightforward and simple way.

Einstein's field equations form a set of ten independent, non-linear, partial differential equations that cannot be solved once and for all with a general scheme. Special solutions can be constructed once certain simplifying assumptions are being imposed. Typically, these come in the form of symmetry assumptions. We thus encounter symmetry considerations a second time at a more specific level. At a fundamental level, they were used to construct plausible physical



theories themselves, for which General Relativity is one example. Now, symmetry assumptions are used a second time to identify classes of solutions of the dynamical equations of this theory. To construct a simple class of cosmological solutions of General Relativity, Friedman first assumed that they should be spatially isotropic and homogeneous. This means that arbitrary spatial rotations and translations should leave the observable universe unchanged for any observer. The solutions so obtained form a class characterised by certain parameters that describe the matter and energy content of the Universe and its expansion behaviour at one point in time. Once these parameters are known or set, the world model is fixed.

The only justification Friedman gave for the symmetry assumptions was mathematical simplicity. Do these assumptions and the world models constructed upon them correspond to reality in any way? As we saw, isotropy demands that the Universe, as seen by any observer, should exhibit the same physical properties in all directions. At first sight, this seems to be manifestly incorrect: The night sky does not at all appear independent of direction. However, if the properties observed in the Universe are averaged over sufficiently large scales, they do in fact approach isotropy. The most striking example for this statement is the cosmic microwave background (CMB), which will be introduced and discussed further below.

If we observe at least approximate isotropy, so should any other observer in the Universe. Since the Copernican revolution, we have grown used to the notion that our location in space and time is by no means unique or central. If, however, the Universe is isotropic about all of its points, as this concept suggests, then it must also be homogeneous.

With these considerations, we have come a long way already. We have seen that theories in physics are constructed upon very general concepts, expressed by symmetries and extremal principles, from which the dynamical differential equations follow. Of the modern theories of physics, only General Relativity is relevant for the construction of world models. In order to find appropriate solutions of its field equations, Friedman introduced the further symmetry assumptions of spatial homogeneity and isotropy, with the sole justification of mathematical simplicity. When combined with these symmetry assumptions, the field equations of General Relativity reduce to the class of Friedman world models. Two questions then arise: First, does our Universe exhibit at least the qualitative features of the Friedman models? Second, if so, is there a unique combination of the cosmological parameters appearing in these models that identifies a single Friedman model out of this class?

## 2 Empirical Evidence for the Standard Cosmology

Let us now investigate the class of Friedman solutions and the two questions raised above. These questions are asking whether the line of reasoning leading from the foundation of physical theories to the construction of physical world models finds its expression in nature. Two aspects of this procedure cannot be overemphasised: First, we have used symmetry assumptions and thus essentially mathematical concepts of regularity and simplicity as guiding principles. The Friedman models are a particularly simple class of solutions of Einstein's field equations. Why should they have any resemblance with the real world we find ourselves in?

The astounding result of decades of cosmological research, as shall be outlined now, is that it is indeed possible to draw a consistent, quantitative picture of our actual universe and its evolution within the class of Friedman models.

Second, while all other areas of physics can conduct experiments with their objects, cosmology cannot. It should be kept in mind that all statements concerning our universe as a whole are based on the comparatively tiny portion of it from which we can receive information. We thus vastly expand the realm of physical laws from our laboratories to the entire observable universe, and we extrapolate from the observable universe to the universe as a whole. The fact that the empirical evidence collected in cosmology does indeed seem to converge with the theoretical concepts underlying the class of Friedman models has a breath-taking aspect.

We shall now go through the most pronounced and relevant empirical pieces of evidence.

1. Friedman models turn out to be generically unstable. They must either contract or expand unless their parameters are very finely tuned. This means that any two points in space identified at a fixed time must either move towards or away from each other not because they move in space, but because space itself drags them along. In expanding models, every observer should see galaxies in his neighbourhood move away from himself with a velocity linearly increasing with distance. The most obvious question to begin with is thus whether our universe is in fact changing with time and whether the galaxies surrounding us do in fact move away from us in the linear fashion that the Friedman models predict.

This is a simple question in principle, but quite hard to address in practice. The problem is that our universe is not ideally homogeneous, as our existence demonstrates. The matter density is not constant, but fluctuates locally. Regions of higher matter density attract neighbouring galaxies and imprint a local motion

on them that is superposed on any motion of cosmic origin. Since any cosmic velocity increases with distance in the Friedman models, the cosmic motion can be expected to dominate the local peculiar motion only beyond a certain distance. Thus, distant galaxies must be precisely observed and their distances measured, which is a demanding procedure. Slipher observed in the 1920s that galaxies typically move away from us, and Hubble found around 1930 that their velocities increase linearly with their distance, just as the Friedman models predict. If this is the correct interpretation of the mean motion of distant galaxies, we seem to be living in an expanding universe. The expansion rate, defined as the relative amount by which cosmological distances increase in time, is called the Hubble constant and is one of the fundamental parameters of any Friedman model.

2. The inverse of the Hubble constant sets the time scale for the cosmic expansion, which turns out to be on the order of 10 billion years. Is this time scale long enough to encompass the observable evolution of the universe, or are there any known objects whose age credibly exceeds the age of the universe? How old are the Earth, the Galaxy and the oldest objects we find in our observable universe?

The decay of suitably long-lived radioactive isotopes such as  $^{235}\text{U}$  or  $^{238}\text{U}$  provides the best constraints of the terrestrial and the galactic ages. The Earth turns out to be 4.6 billion years old. The age of the Galaxy is less well constrained, but likely between 7 and 10 billion years. Older objects exist in the universe whose age we can determine. These are in particular certain end products of stellar evolution, the white dwarfs, and a certain class of co-eval stellar populations, the globular clusters. Upper limits on their age touch approximately 12 billion years. The fact that these age limits broadly agree with the time scale set by the inverse cosmic expansion rate is reassuring.

3. The observation that our universe is expanding today does not necessarily imply that it has been expanding during all of its past. Friedman models which are expanding today but were shrinking or stagnating for part of their history would also be possible. However, a few simple observations show that our universe cannot be of this type. The most intuitive of these is that objects exist whose spectra reveal that the universe was at least six or seven times smaller when their light was emitted than it is today. Thus, if our universe behaves like a Friedman model at all, its present expansion implies that it has always been expanding.

A monotonically expanding Universe keeps shrinking as we go back in time. Any two points then keep approaching each other until they come arbitrarily close after finite time. Any finite section of the Universe must have been very small at

early times. Backward in time, the cosmic matter is compressed by the shrinking volume it is enclosed in. Thus, matter and all other ingredients of the Universe must have been hotter in the past than they are now. If it once was very small, the whole universe may have been as hot as the interior of stars is now. Then, nuclear fusion processes must have occurred throughout the Universe, leading to the formation of light elements such as deuterium, tritium or helium from hydrogen. In fact, interstellar gas contains about 25 % helium and 75 % hydrogen. This large amount of helium cannot have been fused by stars, but only if the entire Universe acted as a nuclear fusion reactor very early during its evolution.

It had been realised by Gamow and his collaborators already in the 1940s that the abundance of helium in the universe can be explained assuming that the Universe itself produced it in a hypothetical hot and dense, early phase. Effective fusion could have set in once the temperature of the Universe had dropped just below a billion degrees, and ended very quickly thereafter as the universe kept expanding and cooling. This happened when the Universe was between two and three minutes old.

4. Charged and sufficiently dense particles at temperatures so high produce energetic thermal radiation. Thus, if the universe was once indeed hot enough to fuse helium, the thermal radiation then produced must still be present, albeit cooled down considerably as the Universe expanded. In fact, it was possible already in the 1940s to predict from the observed helium abundance that the thermal leftover radiation should now have arrived at a temperature of a few degrees Kelvin. Thermal radiation with such a low temperature has characteristic wavelengths in the microwave regime. Thus, the existence of a so-called cosmic microwave background (CMB) could be predicted from the assumption of a hot beginning together with the observable amount of helium. An apparently isotropic, ambient radiation field with properties like the CMB was serendipitously discovered by Penzias and Wilson in 1965 while testing a telecommunication antenna. Immediately, Dicke and co-workers surmised that this radiation could indeed be thermal radiation left over from the very early Universe.

At that time, it was not possible to confirm that the radiation discovered was thermal radiation, as required by this interpretation. However, if one assumed that it was thermal, the temperature corresponding to the measured intensity was found to be approximately 3 Kelvin, in good agreement with the earlier prediction.

A satellite called the Cosmic Background Explorer, COBE in short, impressively demonstrated the thermal nature of the CMB. One of its three instruments measured the electromagnetic spectrum of the CMB and found it to be in perfect agreement with that of thermal radiation with a temperature of 2.7 K.

Another of COBE's instruments solved an acute problem that had accumulated since the discovery of the CMB. Since there are structures like galaxies, galaxy clusters and even larger objects in the Universe, the CMB is not supposed to be ideally homogeneous. Since the present cosmic structures should have originated from predecessors in the very early universe, those should have left their imprint on the CMB. It was estimated that temperature fluctuations with milli-Kelvin amplitudes around the mean temperature should be found. However, when detectors finally reached the required sensitivity, such fluctuations were not detected. Even at a level of one part in a thousand, the CMB was found to be perfectly isotropic.

Since isotropy is one of the primary symmetry assumptions underlying the Friedman models, the remarkable isotropy of the CMB was impressive evidence in their favour. The lack of temperature fluctuations at the level expected from the existing cosmic structures was highly disturbing at the same time. A solution was proposed by Peebles in the 1980s. If cosmic structures consisted not of ordinary matter as we know it, but of a form of matter that does not participate in the electromagnetic interaction, the present cosmic structures could be reconciled with considerably smaller temperature fluctuations in the CMB since then the imprint of the cosmic structures in formation on the CMB could be substantially lower. Fluctuations of one part in 100,000 would then be expected.

At that level, COBE finally found these fluctuations in 1992. This can be seen as a turning point for cosmology, and at the same time as a piece of evidence that cosmic structures are not dominated by the electromagnetically interacting forms of matter that we know, but by some dark matter of hitherto unknown composition.

5. The existence of dark matter was not surprising at that time. Rather, the important result was that dark matter cannot interact electromagnetically. Already in the 1930s, Zwicky had found that the member galaxies of the galaxy cluster in the constellation Coma moved so fast that much more matter was needed to keep them gravitationally bound than could be inferred from the amount of light emitted by the cluster and its galaxies. The amount of mass necessary for balancing the motion of the galaxies was approximately ten times higher than that necessary to produce the light observed. A similar observation was made later at the level of individual galaxies. Their stars also move considerably faster than they should if they moved under the influence of the gravity of their visible matter alone. Dark matter is thus seen on a hierarchy of levels in the Universe, but only the CMB requires it to be of a hitherto unknown form avoiding the electromagnetic interaction.
6. The physics of the CMB and its temperature fluctuations are simple and well understood. The CMB was set free when the universe had become cool

enough for hydrogen atoms to form from the cosmic plasma. It can easily be calculated that the temperature had to drop to approximately 3,000 K for this to happen. When this temperature was reached, the universe was just below 400,000 years old. The hydrogen plasma combined to form hydrogen gas within the relatively short time of about 40,000 years. Since then, the photons of the CMB could propagate almost freely throughout the universe.

Of course, it is not possible to predict the exact structure of the temperature fluctuations in the CMB since they depend on presumably random initial conditions whose exact realisation we cannot know in detail. However, predicting their statistical properties, in particular what the amplitudes of temperature fluctuations of a given size should be like, was possible as early as in 1970. This, however, depends on some of the most important cosmological parameters, such as the densities of ordinary and dark matter, the total matter and energy density in the universe, its expansion rate and the like. The statistical analysis of detailed and sensitive measurements of the CMB temperature fluctuations could thus reveal a good fraction of the cosmological parameters, once compared with theory.

It is an amazing fact on its own that precise measurements of CMB structures confirmed the theoretical predictions in detail and could in turn be used to accurately determine cosmological parameters. For this reason, measuring and interpreting CMB temperature fluctuations has developed into one of the main objectives of current cosmological research. After COBE, two further CMB satellites have been launched. The Wilkinson Microwave Anisotropy probe or WMAP has been observing between 2001 and 2010, while the Planck satellite began operations in 2009. The CMB data taken so far have greatly helped constraining the cosmological parameters with high precision. They have not revolutionised the cosmological model itself, but they were decisive for turning it into the cosmological standard model, whose parameters are now determined typically with relative uncertainties of 10 or less per cent.

7. So far, we have discussed only one piece of evidence probing the late universe, namely the cosmological expansion reflected by the systematic recession of the galaxies in our cosmic neighbourhood. In contrast, the fusion of helium and other light elements and the CMB both probe the early universe, albeit with a large separation in time. Helium fusion ended about three minutes after the beginning, while the CMB was released almost 400,000 years later.

We have touched an important argument that we need to accentuate further: It is possible to interpret these three types of observation in favour of the Friedman models. The recession of the galaxies agrees with the intrinsic instability of the Friedman models and exhibits the expected expansion behaviour. Moreover, it

defines a time scale for the evolution of the universe which agrees reasonably well with the age determinations of old cosmic objects.

Its present expansion suggests that the universe originated in a hot and dense early state, which allowed the fusion of the large amounts of helium that are actually observed. This, in turn, gives rise to the prediction of left-over thermal radiation and thus of the CMB, whose temperature of a few degrees Kelvin is directly related to the amount of helium observed. The level of the temperature fluctuations in the CMB is a strong argument in favour of a form of dark matter that avoids the electromagnetic interaction. The statistics of the CMB temperature fluctuations depend on the details of the cosmic matter content in a precisely predictable way, enabling accurate constraints of cosmological parameters. The abundance of ordinary matter derived from the CMB agrees precisely with the abundance needed to understand the efficiency of the helium fusion.

This indicates that these pieces of evidence do not only individually support the *class* of Friedman models, but that they can be combined to jointly support a *single* Friedman model. This is an important step forward. Friedman models do not only allow the interpretation of snapshots of the universe taken at vastly different times, but they seem to single out one specific Friedman model that allows the consistent interpretation of all cosmological evidence discussed so far.

8. This picture can be extended by a few more colourful strokes. Further evidence is available that probes the Friedman models at epochs intermediate between the CMB and today.

Exciting and lively debated is the direct measurement of the cosmic expansion by means of a particular type of stellar explosion, the so-called supernovae of type Ia. We believe that such explosions arise when a white dwarf star is driven above its upper mass limit by matter overflowing from a companion star. Above this well-defined mass limit, the white-dwarf material is explosively ignited which disrupts the entire star. The amount of exploding material is thus known, approximately 1.4 solar masses, and therefore also the energy released, which sets the luminosity of the supernova. From the observed flux, we can then infer its distance. Its spectrum reveals when in cosmic history the supernova exploded. Type-Ia supernovae thus allow the reconstruction of the evolving of distances with the cosmic expansion, i.e. they directly probe the cosmic expansion history.

In doing so, they reveal an astonishing fact: When the universe was about half as old as it is now, its expansion began to accelerate. This is utterly counter-intuitive. We expect gravity to decelerate the cosmic expansion because of the usual gravitational attraction. Accelerated expansion is allowed, however, by General Relativity, provided there is a substance that Einstein introduced under

the name of cosmological term or cosmological constant in order to stabilise the intrinsically unstable Friedman models.

We do not know what the cosmological constant could be. Attempts at explaining it in terms of a quantum field lead to the concept of dark energy, introduced for the sole purpose of interpreting the accelerated cosmic expansion indicated by type-Ia supernovae.

Strange as they may sound, these ideas receive substantial support from the CMB. Among the most solid conclusions from the statistical analysis of the CMB temperature fluctuations is the insight that the universe must be spatially flat. This is concluded directly from the size of the most pronounced warm and cool spots in the CMB. At fixed physical size, they appear larger if space is positively curved, and smaller if it is negatively curved. However, spatial flatness in the Friedman models is possible only if the matter or energy densities of all components of the cosmic fluid add up to a critical value of about one proton in five cubic metres. We thus know what the total matter and energy density in the universe is, but we know also what the densities of dark and ordinary matter are. Both together sum up to only about 30 % of the known total amount. If the difference is contributed by the cosmological constant or the dark energy, a model emerges which can precisely reproduce the expansion history probed by the type-Ia supernovae.

Another important class of observations probes the large-scale structure in the universe. Galaxies are not randomly distributed in space. Rather, they form galaxy clusters and extended filamentary structures, many millions of light-years long. Like the structures in the CMB, these structures in the distribution of cosmic objects carry most valuable statistical information. In particular, there is a characteristic length scale imprinted into the galaxy distribution which was set at a very specific epoch in cosmic history which is defined by the total matter density compared to the energy density of radiation in the universe. This implies that, if this characteristic scale in the galaxy distribution can be measured, the matter density can be inferred from it.

This approach requires galaxy surveys extending to distances that are substantially larger than the characteristic scale to be measured. Surveys of such size have become possible only in the recent past. They confirmed that the total matter density is about 30 % of the critical value, in agreement with the CMB data.

Yet another probe of cosmic structures dominated by dark matter is provided by the so-called gravitational lensing effect. General Relativity implies that concentrations of mass or energy deflect light in a way comparable to convex optical lenses. This gives rise to a multitude of interesting effects of different magnitude. In our context, the most important one is that any light ray propagating from a distant source to us must be deflected multiple times by the intervening large-scale structures, irrespective of what kind of matter they are composed of. This



deflection gives rise to faint distortions of background galaxies which are indeed measurable, albeit with a formidable effort.

This cosmological weak lensing effect cannot distinguish between diluted matter that is clumped to a large degree and dense matter that is less clumpy because it is only sensitive to the absolute amount of inhomogeneity in the matter distribution. However, the results are well in agreement with a Friedman model in which the matter density reaches approximately 30 % of the total, critical density, while the rest is contributed by the cosmological constant or dark energy.

### 3 Consequences and Perspectives

What does it all mean? We have collected a substantial body of evidence in favour of the Friedman class of cosmological models. It is worth recalling what they are based upon: The only two ingredients were General Relativity, combined with spatial isotropy and homogeneity. Going one level deeper, General Relativity itself is built upon the concepts that the geometry of space-time adapts to the presence of matter and energy and that the experimentally well-established theory of Special Relativity remains locally valid. The dynamical equations governing the way how geometry reacts to the presence of matter and energy again follow from underlying symmetry and extremal principles extending far beyond General Relativity itself. Interestingly, it can be mathematically proven that General Relativity is unique in a quite general sense. Under broadly acceptable assumptions, Einstein's field equations are even the only dynamical equations possible.

The class of Friedman world models thus seems to stand on a rock-solid theoretical foundation, supported by a large body of empirical evidence. However, ways out are possible along the paths sketched in the beginning.

Either, one remains within General Relativity, then at least one of the two symmetry assumptions must be abandoned that the Friedman models are built upon. Any deviations from symmetry must, however, obey the tight limits on isotropy set in particular by the temperature of the CMB and its fluctuations. More vulnerable is the assumption of homogeneity, which is much harder testable, if at all. If we decided to give it up, we would have to accept being located not at a random, but at a fairly special place in the universe. While this is not at all impossible, it is quite unlikely that a sufficiently special place exists from where the universe looks as peculiar as it does, in particular in view of its accelerated expansion.

Alternatively, we could give up or modify General Relativity. The most gentle way of doing so consists in adding terms to the action that are still in agreement with the general underlying symmetries. The principle of least action then pro-

vides a standardised way of deriving modified field equations to replace Einstein's equations. Based upon them, Friedman's symmetry assumptions could be re-established to arrive at modified or generalised Friedman models. Alternatively, at least isotropy could be questioned in addition. However, General Relativity has so far survived all experimental tests it was subjected to. Admittedly, the most stringent tests all concerned local, weak gravity in the Solar System, but nonetheless these must also be met by alternative theories.

More radical approaches are also possible and are being pursued. One consists in extending General Relativity to more than four space-time dimensions. This was already suggested by Klein and Kaluza in the 1920s in attempts to unify electromagnetism and gravity and to explain quantum aspects of matter. The additional, fifth, dimension then introduced had to be considered as compactified, or rolled up, in order to be macroscopically hidden. This concept has been revived in current theories. Another approach aims at a quantum theory of gravity, which still seems well beyond the horizon.

Perhaps the most conservative point of view accepts that the foundations of the Friedman models are hard to shatter. Then, accepting the Friedman models and testing them against the empirical evidence leads to the single, standard model of cosmology that provides a consistent framework for virtually all cosmological observations. It comes, however, at the considerable price that dark matter and dark energy must then be accepted. We have some promising and testable ideas regarding the nature of the dark matter. Most likely, it is composed of weakly interacting, massive elementary particles. No suitable particle has yet been discovered, but it seems plausible that if dark-matter particles exist, either indirect evidence for them will be found at the Large Hadron Collider, or direct evidence in dedicated recoil experiments.

Dark energy, in contrast, remains essentially mysterious to us. It could be Einstein's cosmological constant, but then its theoretical foundation seems unsatisfactory. It could be some quantum field taking part in the cosmological evolution in some way, but so far there is no empirical evidence whatsoever that the dark energy might depend on time. It is well possible that, if General Relativity persists, we have to accept the cosmological constant in just the same way as we have to accept other constants of nature, such as the fine-structure constant or the elementary charge.

Unveiling the nature of the dark matter and the dark energy are at the heart of current cosmological research. In the context of cosmological model building, this is perhaps an irrelevant detail. What is important, however, is a remarkable reversal in the order of arguments usually leading to the construction or the dismissal of a model. The foundations of the Friedman class of cosmological models appear so solid that it seems more appropriate to accept the seemingly exotic

consequences of dark matter and dark energy than to abandon the model. The situation reminds of a letter Einstein wrote to Sommerfeld after he had completed General Relativity. In this letter, Einstein remarked that he would lose no words in defending the theory because Sommerfeld would be convinced of it at a glance. In the cosmological standard model, the simplicity and the high degree of symmetry of the primary assumptions seem more appealing than the apparently preposterous consequences might be repelling.

## 4 Dark matter, dark energy, and the future of the Universe

Up to this point, the cosmological standard model may appear impressive by its simplicity and its consistency throughout almost all of cosmic history. We have, however, swept one of its major problems under the rug, which has to do with an apparent violation of causality.

As we have seen before, the CMB was released from the cosmic plasma when the universe was approximately 400,000 years old. During this time, light can obviously travel by no more than 400,000 light years. When compared to the full CMB sky, this distance corresponds to a very small angular scale. It spans an angle approximately as large as twice the full moon. Since no information can propagate faster than the speed of light, two points on the CMB separated by more than twice the so-called horizon radius of 400,000 light years could never communicate prior to the release of the CMB. How was it possible then that any two points on the CMB separated by more than a few angular diameters of the full moon could ever have arranged to attain the same temperature? How could the temperature information at one point on the CMB sky ever have propagated far enough to adjust the temperature to the same value everywhere?

One might object that this is of course necessarily so – in a model universe that has been set up to be ideally isotropic. By construction, the temperature must then be the same everywhere on the observer's sky, thus the observation of a CMB sky with constant temperature just appears as a consequence of the far-reaching symmetry assumptions that we started out with. This is not a way out, however, because coherent structures exist in the CMB that are larger than the horizon radius. Thus, even if we would be willing to accept isotropy of the CMB temperature without asking further how it could have been established in absence of causal mechanisms, the existence of coherent structures larger than the horizon radius implies that the processes creating these structures must have

acted in a causally connected way even though the structures extend well beyond the scale of causal connection. This is an unbearable imposition.

The only feasible way out seems to be postulating a very early epoch in the cosmic evolution in which the universe expanded in a very strongly accelerated fashion. This epoch is called cosmological inflation. Among the primary purposes of its introduction was the causality problem just sketched. It solves this problem by assuming that tiny, causally connected regions in the primordial universe were stretched to cosmological size by the inflationary expansion, turning a potentially small section of a region that was previously in causal contact into our observable universe.

It is unclear what this epoch of cosmological inflation could have been driven by. A suitable quantum field called the inflaton is postulated for this purpose. This may seem helplessly unsatisfactory, but it has observable consequences. One of them is that any quantum field must undergo fluctuations because of Heisenberg's uncertainty principle. During the inflationary epoch, these quantum fluctuations would have been stretched to macroscopic and even cosmological scales such that they could later form the seeds for the rich variety of cosmic structures we see today. Even though it may appear ludicrous, this hypothesis allows a calculation of the expected statistical properties of cosmic structures produced that way. Measurements of the temperature fluctuations in the CMB confirm this expectation precisely.

This gives rise to the truly breath-taking notion that cosmic structures may have originated in quantum fluctuations of a primordial inflaton field that drove the early phase of accelerated expansion. If this seems incredible, we must recall that we have now entered another phase of accelerated cosmic expansion, as demonstrated directly by the type-Ia supernovae and indirectly by the CMB.

## References

Singh, Simon (2005). *Big Bang. The Origin of the Universe*. New York: Harper Perennial.

**Prof. Dr. Matthias Bartelmann**

Center for Astronomy of the University of Heidelberg

Institute of Theoretical Astrophysics

Philosophenweg 12

69120 Heidelberg

Germany

bartelmann@uni-heidelberg.de

Andreas Bartels

# The Standard Model of Cosmology as a Tool for Interpretation and Discovery

Commentary on Matthias Bartelmann

Science does not live with facts alone. In addition to facts, it needs models. Scientific models fulfill two main functions with respect to empirical facts. First, they provide a net of theoretical relations by which we may *interpret* the data. By embedding data into the standard model of cosmology (sometimes by identifying the data as fulfilling a certain *prediction* of the theory), a fact about the universe that would otherwise be rather contingent and unrelated to other facts, will be located at a particular place in the causal net of the model, and hence will be supplied with *evidential* status with respect to other parts of the model. In reverse direction, the data thus integrated in the model may fulfill *evaluative* function: When further analyzed, they turn out to *confirm* or to *disconfirm* theoretical relations of the model. In the latter case, the model has to be modified or to be rejected altogether. Empirical data, which have been integrated into the model, cannot only confirm or disconfirm the model, but they can also be used to *specify* the values of theoretical parameters of the model.

This sort of interaction between the data and the model constitutes what may be called the *descriptive* (or puzzle solving) dimension of science. But there is also a *explorative* function of models that is responsible for the research dynamics of cosmology. Models do not only *describe* reality, they are also instruments for *exploring* reality. They are not only involved in the *integration* of known data, but also in the *discovery* of new data. This function is demonstrated by the standard model's prediction of the existence of dark matter and dark energy in the actual universe. The model requires, under the assumption of some previously accepted interpretations (flatness and critical density), that facts, not yet detected by observations, do exist. In order to be able to play that sort of role, the model must have gained high reputation (with respect to former successes in integrating data). Because of this reputation (contrary to the case of testing the model), not the model, but the observations are blamed for the disagreement of the observed data with predictions of the model. In contrast, for example, with the case of the deviation of the orbit of Uranus, compared to the predictions of Newtonian gravitation theory, assuming that some additional boundary condition is present that has not yet been detected by observation, would be no option. All possible boundary conditions have already been included in the model. The fact is also not

conceived as an anomaly of the theory, but on the contrary, as indicating that, for the actual observations, there are *some hidden facts that have not been uncovered by these observations*. Actually, the model shows that our observations have been blind for important facts as yet; facts required by the authority of the model – this makes the case distinct to the usual case of prediction of novel facts by a model.

Matthias Bartelmann's presentation of the present state of research in cosmology nicely demonstrates how these four different sorts of interactions between models and facts are actually fulfilled: Interpretation (of observed facts by the model), evaluation (of the model by the facts), specification (of the model by the facts), and exploration (of not yet observed facts by the model).

Interpretation: The discovery of the cosmic microwave background (CMB) exemplifies a case of interpretation. After the unintentional finding of the isotropic radiation background by Wilson and Penzias in 1965, it appeared that this radiation might be interpreted as the microwave background predicted by Gamow and collaborators in the 1940s as the relict of the hot period of the universe near the big bang, in which the helium observed in the actual universe has been produced. The general background that made this interpretation possible was the Friedman class of models as a tool for the scientific understanding of the actual universe. Questions that had to be answered in order to launch that interpretation were: First, is the measured radiation actually *thermal* radiation, as the model requires? (This question has been answered to the positive by the later COBE findings). Second, are there fluctuations in the radiation as they have to be expected in order to account for the observed inhomogeneity of the matter distribution of the observed universe? To this question, the researchers, in the first instance, did not get the expected answer: the necessary fluctuations did simply not appear.

At that point the interpretation relation between the data (the COBE measurement results) and the model turns into an explorative relation: The disagreement between the measurements and the alleged inhomogeneity could possibly be removed, if the model would be enriched by additional mechanisms. Such an additional mechanism is the invention of the dark matter hypothesis: "If cosmic structures consisted not of ordinary matter as we know it, but of a form of matter that does not participate in the electromagnetic interaction, the present cosmic structures could be reconciled with considerably smaller temperature fluctuations in the CMB since then the imprint of the cosmic structures in formation on the CMB could be substantially lower" (Bartelmann 2011, 6). According to the dark matter hypothesis, dark matter had no interaction with light, and thus does not show up in the CMB data in the way ordinary matter does. If no such additional mechanism would have been available to remove the disagreements of the CMB data with the theoretical requirements, then this would have meant a potential negative evidence for the standard model of cosmology.

**Evaluation:** Actually, the measured temperature fluctuations in the CMB data have turned out to be the most decisive tool for testing the standard model. First, in 1992, the new fluctuation predictions, on the basis of the dark matter hypothesis, were confirmed by COBE measurements. Bartelmann comments this finding as the decisive breakthrough in recent cosmology: “This can be seen as a turning point for cosmology, and at the same time as a piece of evidence that cosmic structures are dominated not by the electromagnetically interacting forms of matter that we know, but by some dark matter of hitherto unknown composition” (Bartelmann 2011, 6). What appeared as a major challenge for the original standard model in the first instance, had turned out to provide some impressive confirmation of the enlarged version of the model. Even if the CMB data gained their theoretical relevance only by means of the background of an interpretation provided by the standard model, and thus were “theory-dependent” in that sense, this did not result in a problematic status of that empirical data concerning their capacity for testing the standard model. Instead, it appeared that the CMB data entail aspects conflicting with the original standard model and thus provoke a modification of it. Again, the prediction’s compatibility resulting from that modification with the CMB data was not self-guaranteed, but actually appeared. In retrospect, CMB not only provided positive evidence for the standard model, but also turned out to work as a detector for the limits of the model – a circumstance that increases the empirical credibility of the model all the more. Since the model’s confirming evidence not simply “fits” the model’s predictions, but discloses some missing pieces of the puzzle, the evidence turns out to be highly *model-independent*. There obviously exists no self-contained agreement between the data and the model produced by the model’s interpretation of the data.

**Specification:** Empirical data do not only have the capacity to confirm or to disconfirm a model. They can also be used to *specify* the values of the model’s theoretical parameters. This connection between confirmation and specification of theoretical parameters has most clearly been pointed out by Clark Glymour in his bootstrap model of confirmation (Glymour 1980). He claimed that in interesting cases of theory confirmation the confirming evidence will not speak in favor of a theory, unless the evidence has been used, in connection with some part of the theory, as a resource to specify values of some parameters of the theory. To speak in favor of a theory therefore means to specify the values of those theoretical parameters that must be known to the scientist in order to enable him to determine whether the data satisfy some equations of the theory.

There are strong theoretical connections, according to the enlarged standard model, between the statistical properties of the temperature fluctuations in the CMB and important cosmological parameters, such as the densities of ordinary and dark matter, the total matter and energy density, and the expansion rate of

the universe. Thus, the measurement results for these statistical properties also allow for specifications of those important parameters (cf. Bartelmann 2011, 7). Furthermore, the statistical analysis of the CMB, provides a measure of the global curvature of space. The results have been in favor of the global flatness of the universe; this, in turn, means that the universe must have a *critical density* which is about three times the observed matter density of the universe. The determination of the critical density leads to a prominent further prediction of the model: Even if dark matter is included, there is a gap of 70% between the critical density and the matter-energy of the universe. This means that there must be a high amount of *dark energy* in the universe.

Exploration: I have already mentioned that the use of the CMB data for the evaluation of the standard model also discloses some theoretically highly relevant disagreements between the data and the model. The data roughly fit the model, but to yield a precise fit, some modifications of the model would be required. Thus, the example of the recent development of the standard model of cosmology confirms the insight proposed by philosophers of science, such as Kuhn and Lakatos, according to which a disagreement between data and the model does not necessarily lead to a “falsification” of the model. There is an alternative scenario of postponed falsification, according to which the scientists continue using the model as long as all possible resources to remove the disagreement by modifications, either in initial and boundary conditions or in the equations of the theory, have been exhausted. But even if this is not plainly wrong, it seems at least to miss a decisive point as highlighted by our current example: Depending on the credibility of the corresponding model – and the credibility of the standard model of cosmology was extremely high when CMB was discovered – it can happen that the scientists take the disagreement as a positive indication of some not yet detected mechanisms. Then the data – in connection with the model – no longer figure as a tool for confirming or disconfirming an existing theory, but seem to transmute into a tool of *exploration* aiming at the discovery of possible missing pieces in the theoretical picture. Confirmation and disconfirmation are present also then, but related to the testing of particular mechanisms the introduction of which have been provoked by the data.

The new pieces within the theoretical picture, as demonstrated by the case of the cosmological standard model, may exhibit new theoretical riddles. Such a riddle is presented by Bartelmann in the last part of his paper, titled “Dark matter, dark energy, and the future of the Universe”. The high isotropy of CMB in the recent period of the universe, Bartelmann argues, provokes the question of “how could the temperature information at one point of the CMB sky ever have propagated far enough to adjust the temperature to the same value everywhere” (Bartelmann 2011, 12). This again is the starting point for the invention of a new



mechanism to be added to the standard model, namely the mechanism of the *inflationary expansion*.

The example of the standard model of cosmology clearly demonstrates how methodology of science misses scientific practice, if it is only devoted to the confirmation-or-falsification aspect of the dynamics of theories. New data drive the development of a theory not only by providing new information with respect to the evaluation of theories, but also – and sometimes more importantly – by provoking new ideas concerning the incorporation of new mechanisms into the theory. The observations not only pass their judgment on the predictions of a model – sometimes they drive the generation of new predictions. In those cases, the observations – on the basis of a well-established model – initiate the discovery of real mechanisms that the former picture had neglected.

## References

- Bartelmann, M. (2013). *Cosmology – The largest possible model?* This volume.  
 Glymour, C. (1980). *Theory and Evidence*, Princeton: Princeton University Press.

### **Prof. Dr. Andreas Bartels**

University of Bonn  
 Institute of Philosophy  
 Am Hof 1  
 53113 Bonn  
 Germany  
[andreas.bartels@uni-bonn.de](mailto:andreas.bartels@uni-bonn.de)



Martin Golubitsky

## Patterns in Physical and Biological Systems

Mathematics can be applied in many ways in science, but let's begin by focusing on one typical caricature. Study an application until it is possible to derive a detailed mathematical model. Then use mathematics (by which we include both analysis and computation) to solve that model and make predictions. Compare the results of the model with experiments; if there is a discrepancy refine the model and iterate the process. Spectacularly successful examples of this caricature include the  $n$ -body problem (a model for planetary motion) and the Navier-Stokes equations (a model for fluid motion) – though there are many other examples.

The question that we want to discuss here is what happens when a model is too complicated to be analyzed or when no detailed model can be derived. Can mathematics still be used to help understand that application and even to make predictions? The answer is yes – but one must ask the right kind of question.

The common approach is to understand the structure that a detailed model must have and then use that structure to make predictions about the kinds of solutions one can expect the unknown equations to produce. In the past 50 years this meta-principle has appeared in a number of different guises including, for example, catastrophe theory (R. Thom, 1972; E.C. Zeeman, 1977), bifurcation theory (J. Guckenheimer and P. Holmes, 1983; M. Golubitsky and D.G. Schaeffer, 1985), and symmetry-breaking and pattern formation (L. Michel, 1972; D.H. Sattinger, 1979; M. Golubitsky et al., 1988; M. Golubitsky and I. Stewart, 2002). In these theories some structure is assumed and then the kinds of solutions consistent with that structure are classified. Also, in these theories new solutions are found by classifying typical transitions as parameters are varied.

For example, catastrophe theory classifies the expected transitions between critical points as parameters are varied (assuming that the model has a potential function) and bifurcation theory classifies the expected kinds of dynamics that occur in systems of differential equations near an equilibrium that loses stability as a parameter is varied. In both theories the expected transitions depend on the number of (independent) parameters that the model is assumed to have. In symmetry-breaking and often in pattern formation the additional assumed structure is a group of symmetries for the model equations.

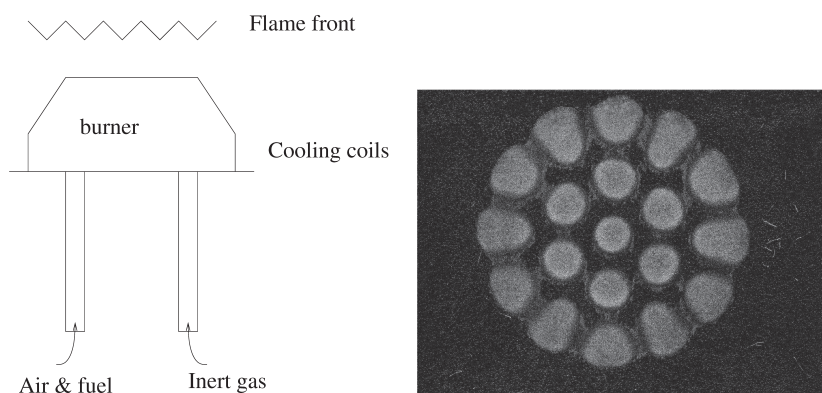
This article will focus on symmetry-breaking and pattern formation in its simplest form. We will discuss two applications where no detailed system of model equations is known, but where a group of symmetries for these unknown equations can safely be assumed. We will assume that there is a homogeneous (or

group invariant) equilibrium and classify the symmetry properties of new solutions when that equilibrium loses stability (a symmetry-breaking bifurcation) as a single parameter is varied. And then – we will interpret these results for the application. The focus will be on applications and predictions; only references will be given for the needed mathematics. Our exposition will follow closely the descriptions of these applications given in *The Symmetry Perspective* by (M. Golubitsky and I. Stewart, 2002) (indeed some of the material is taken verbatim from this volume). This reference also supplies many of the mathematical details behind the arguments that we give here.

## 1 Patterns in Flames

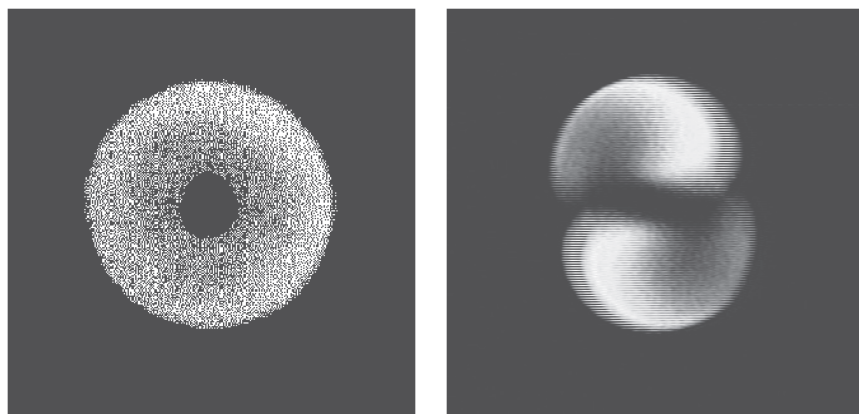
There is a huge literature on patterns in a variety of classical fluid dynamical and chemical reacting systems including the Taylor-Couette experiment, Bénard convection, the Faraday experiment, and the Belousov-Zhabotinskii reaction. See, for example, the references in (M. Golubitsky and I. Stewart, 2002). An experimental system that has received somewhat less discussion is the pattern-rich porous plug burner studied for many years by the physicist Michael Gorman at the University of Houston (M. Gorman et al., 1994a,b).

A cross-section of Gorman's system is shown in Figure 1 (left). Viewed from above the burner is circularly symmetric. The flame is ignited on top of the burner and maintained by the fuel flowing continuously through the burner. A typical steady flame pattern is also shown in Figure 1 (right).

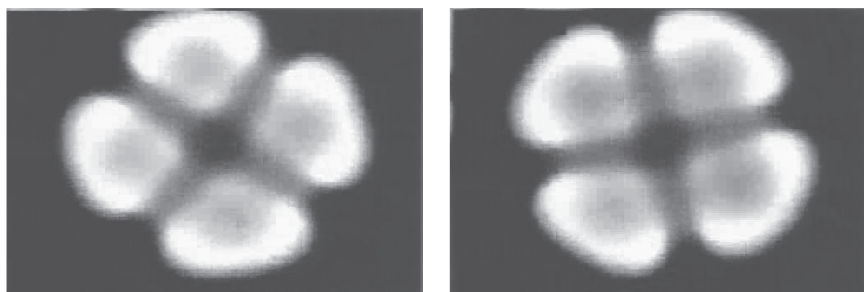


**Figure 1:** Cross section of a porous plug burner and a typical pattern formed by the flame on the burner's top surface. Images courtesy of M. Gorman.

Symmetry enters the discussion of flames most prominently through time-periodic states. A theorem that has been proved many times in the literature in specific applications, but whose validity depends only on the existence of circular symmetry (M. Golubitsky et al., 1988), is the following. When a circularly symmetric equilibrium of a circularly symmetric system (see Figure 2 (left)) loses stability to time-periodic oscillations two states form: rotating waves and standing waves. A *rotating wave* is a state whereby time evolution of the state is given by rigid rotation and a *standing wave* is a time-periodic state that has at least one line of symmetry for all time. The physical implication is that when a rotating wave is found in an experiment, it can be presumed that standing waves are also present; hence it is not surprising that the standing waves will also be observed.



**Figure 2:** Flames on circular burner. (Left) Circularly symmetric flame; (right) rotating two-cell flame. Images courtesy of M. Gorman.



**Figure 3:** Standing wave flames on circular burner. Two images on one trajectory illustrating same four lines of symmetry. Images courtesy of M. Gorman.

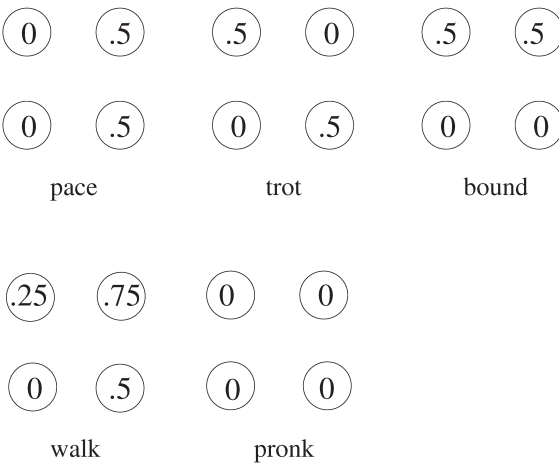
This is precisely what Gorman found. He observed a rotating wave in the flame experiment (see Figure 2 (right)) and sometime later (a year or so, as it happened) Gorman also found the standing wave (see Figure 3).

## 2 Quadruiped Central Pattern Generators

It is well known that all horses walk and that some horses trot while others pace. In addition squirrels bound and deer will sometimes prong. There is one feature that is common to all gaits: they are repetitive; that is, they are time-periodic.

In the pace, trot, and bound the animal's legs can be divided into two pairs – the legs in each pair move in synchrony, while legs in different pairs move with a half-period phase shift. The two pairs in a *bound* consist of the fore legs and the hind legs; the two pairs in a *pace* consist of the left legs and the right legs; and the two pairs in a *trot* consist of the the two diagonal pairs of legs. The quadruiped *walk* has a more complicated cadence (each leg moves independently with a quarter-period phase-shift in the order left hind, left fore, right hind, and right fore), whereas the quadruiped *prong* is a simple motion (all four legs move synchronously).

We summarize the descriptions of these five gaits in Figure 4 by indicating the phases in the gait cycle when each given leg hits the ground. For definiteness, we start the gait cycle when the left hind leg hits the ground.



**Figure 4:** Five standard quadruipedal gaits.

(J.J. Collins and I. Stewart, 1993, 1994) and (G. Schöner et al., 1990) made the observation that each of these gaits can be distinguished by symmetry in the following sense. Spatio-temporal symmetries are permutations of the legs coupled with time shifts. So interchanging the two fore legs and the two hind legs of a bounding animal does not change the gait, while interchanging the two left legs and the two right legs leads to a half-period phase shift. In a walk permuting the legs in the order left hind to left fore to right hind to right fore leads to a quarter-period phase shift. Based on these gaits we consider three symmetries: the bilateral symmetry that simultaneously interchanges left legs and right legs; the transposition that interchanges front and back legs; and the walk symmetry. Table 1 lists which of these symmetries are applicable to each gait and, if applicable, the associated phase shift.

**Table 1:** Phase shifts corresponding to leg permutation symmetries in standard quadrupel gaits

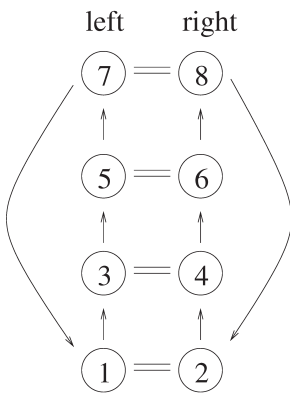
Gait	left-right	front-back	walk
trot	1/2	1/2	n.a.
pace	1/2	0	n.a.
walk	1/2	n.a.	1/4
bound	0	1/2	1/2
pronk	0	0	0

Biologists often make the assumption that somewhere in the nervous system is a locomotor *central pattern generator* or CPG that produces the rhythms associated to each gait. CPGs are known to exist in primitive animals but they have not been identified in mammals. Nevertheless, suppose we assume that there is a locomotor CPG in quadrupeds – how can we model it? Neurons themselves are modeled by systems of differential equations (for example, the Hodgkin-Huxley equations (J. Keener and J. Sneyd, 1998)) and CPGs are thought to be a coupled array of neurons (see (N. Kopell and G.B. Ermentrout, 1988, 1990), (G. Schöner et al., 1990), (R.H. Rand et al., 1988)). So we may assume that our model is (a perhaps large dimensional) system of coupled ODEs. What structure may we assume that such a system of equations should have?

We imagine that for each leg there is a single group of neurons whose job is to signal that leg to move, and that the groups of neurons are otherwise identical. Moreover, we assume that the groups of neurons are coupled in some manner – and to simplify matters we assume that the kinds of coupling fall into a small number of identical types. A natural mathematical question now arises – even at this level of generality. Can couplings between these four groups of neurons be

set up so that periodic solutions having the rhythms associated with each of these gaits exist? The answer is, perhaps surprisingly, no. The reason for this is subtle. It is known that trot and pace are different gaits. However, if a four group system were capable of producing periodic solutions with the symmetries of walk, trot, and pace, then walk and trot must be the same up to symmetry and would for all practical purposes be the same gait.

The next simplest model would have eight groups of neurons with each leg receiving signals from two different groups of neurons. (M. Golubitsky et al., 1998) introduced the network shown in Figure 5 by assuming that the eight-node network should independently have both bilateral  $\kappa$  symmetry and the four-cycle walk symmetry  $\omega$ . Thus the symmetry group of the eight-cell quadruped CPG is  $\Gamma = \mathbf{Z}_2(\kappa) \times \mathbf{Z}_4(\omega)$ . For expository purposes we assume that cells 1, ..., 4 determine the timing of leg movements, and refer to the remaining four cells as 'hidden'. We also follow (M. Golubitsky et al., 1999) and show how the mathematical analysis of the structure of this CPG network can still lead to testable predictions about the structure of gaits.



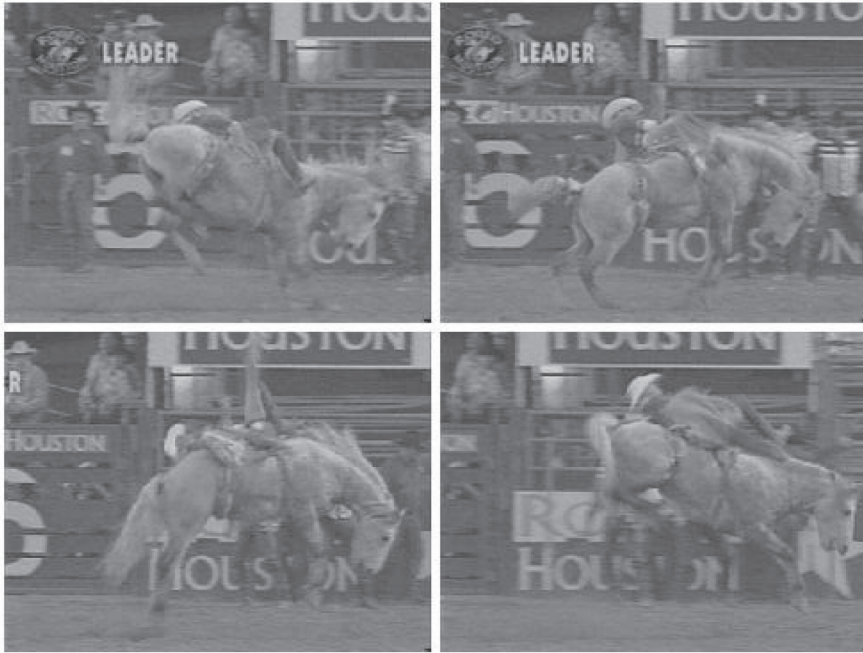
**Figure 5:** Eight-cell network for quadrupeds. Double lines indicate contralateral coupling; single lines indicate ipsilateral coupling. Direction of ipsilateral coupling is indicated by arrows; contralateral coupling is bidirectional.

In fact, the eight-cell network in Figure 5 (right) is essentially the only one that can produce periodic solutions with the spatio-temporal symmetries of walk, trot and pace (M. Golubitsky et al., 1998, 1999; P.L. Buono and M. Golubitsky, 2001). Next we ask the question: Which periodic solution types can be expected to emanate from a stand equilibrium in systems of differential equations associ-



ated with this cell network. We call these gait types *primary* gaits. It turns out that such systems can produce a non-standard gait in addition to the five gaits we have discussed previously. This gait is called the *jump* and can be described as ‘fore feet hit ground, then hind feet hit ground after one beat, then three beats later fore feet hit ground’. The existence of this quadruped gait is a prediction of the model.

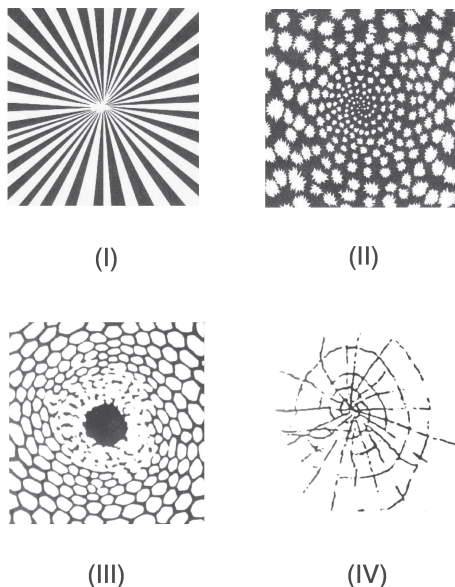
Indeed, we observed a gait with that spatio-temporal pattern of the jump at the Houston Livestock Show and Rodeo. Figure 6 shows four equal time-interval video frames of a bucking bronco. The timing of the footfalls is close to 0 and  $1/4$  of the period of this rhythmic motion. Later on we found that (P.P. Gambaryan, 1974) had identified the *primitive ricocheting jump* of a Norway rat and an Asia Minor gerbil that also has the cadence of the jump.



**Figure 6:** Approximate quarter cycles of bareback bronc jump at Houston Livestock Show and Rodeo.

### 3 Geometric Visual Hallucinations

(H. Klüver, 1966) observed that geometric visual hallucinations divide into four *form constants*: tunnels and funnels; spirals; lattices including honeycombs and phosphenes; and cobwebs. See Figure 7. (P.C. Bressloff et al., 2001, 2002) are able to explain the origins of the four form constants as symmetry-breaking with respect to the Euclidean group of planar translations, rotations and reflections as it acts on the primary visual cortex (V1). In this section we will describe that action.



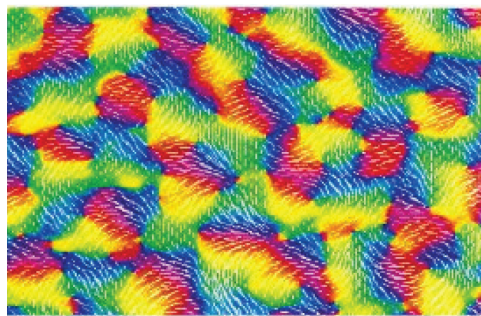
**Figure 7:** Hallucinatory form constants. (I) funnel and (II) spiral images seen following ingestion of LSD (R.K. Siegel and M.E. Jarvik, 1975), (III) honeycomb generated by marihuana (J. Clottes and D. Lewis-Williams, 1998), (IV) cobweb petroglyph (A. Patterson, 1992).

The idea of viewing the origin of geometric visual hallucinations dates to the work of (G.B. Ermentrout and J.D. Cowan, 1979). Ermentrout and Cowan argue that when an individual is under the influence of a drug, the entire primary visual cortex is stimulated uniformly by the drug and not by the retina. When this forced stimulus is sufficiently large, patterns of activation are formed on V1 and interpreted by the brain as visual images – often with a distinctly geometric flavor. However, the work in (G.B. Ermentrout and J.D. Cowan, 1979) was completed

before the nature of coupling of neurons in V1 was understood. Thus (G.B. Ermentrout and J.D. Cowan, 1979) assumed that models of V1 are Euclidean-invariant with respect to the standard action of the Euclidean group on the plane and symmetry-breaking arguments only led to two of the four form constants (funnels and spirals).

In this section we present part of the discussion of V1 in (M. Golubitsky and I. Stewart, 2002) (much of it verbatim), which itself is an abbreviated version of the discussion in (P.C. Bressloff et al., 2001). In mammalian vision, neurons in V1 are known to be sensitive to the orientation of contours in the visual field. Moreover, as discussed in (P.C. Bressloff et al., 2001), the pattern of neuronal connections in V1 leads to a specific action of the Euclidean group that is different from the standard one on the plane.

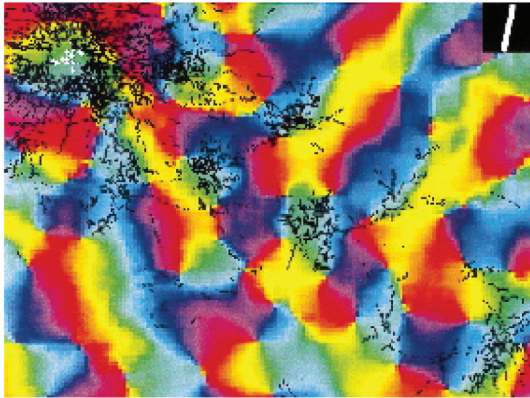
The V1 layer is approximately a square, 40mm on a side. (D.H. Hubel and T.N. Wiesel, 1974a,b,c) noted that V1 is divided into small areas of about 1mm diameter, called *hypercolumns*, and the neurons in each hypercolumn receive signals from one small area in the retina. A hypercolumn contains all cortical cells that correspond to such an area: its architecture allows it to determine whether a contour occurs at that point in the retinal image, and if so, what its orientation is. This task is accomplished by having all pairs of cells in a hypercolumn connected by inhibitory coupling – so if a contour is detected by one neuron, it tends to suppress the other neurons in that hypercolumn, a local *winner-take-all* strategy. Experimental confirmation of the existence of hypercolumns is found in (G. G. Blasdel, 1992), see the iso-orientation patches in Figure 8.



**Figure 8:** Distribution of orientation preferences in V1 obtained via optical imaging. Redrawn from (G. G. Blasdel, 1992).

What is curious – and crucial from the symmetry point of view – is how hypercolumns themselves are coupled. In recent years information has been obtained

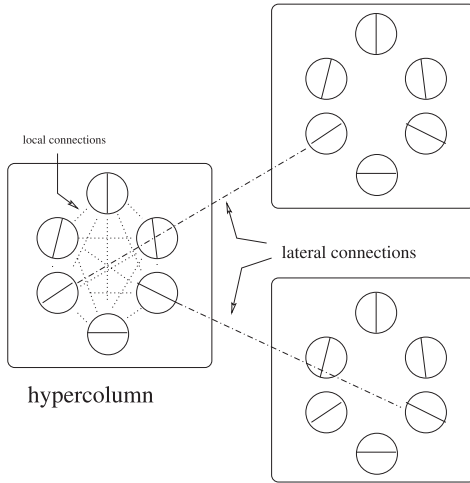
about connections using, for example, optical imaging with voltage-sensitive dyes (W.H. Bosking et al., 1997). These studies show that cells that selectively fire for one orientation make contact only every millimeter or so along their axons with cells that fire selectively in the same orientation. See Figure 9, which illustrates the inhomogeneity in lateral coupling.



**Figure 9:** Lateral connections made by a cell in V1 superimposed on iso-orientation patches. Redrawn from (W.H. Bosking et al., 1997).

In addition, it appears that the long axons that support such connections, known as *intrinsic lateral* or horizontal connections, tend to be oriented more or less along the direction of their cells' preference. See the schematic diagram in Figure 10. Note that the strength of the lateral connection between hypercolumns is small when compared to the strength of the local connections within hypercolumns. These observations lead to the schematic pattern of neuronal connections shown in Figure 10.

Observe that when one makes the hypercolumns infinitesimal then the resulting schematic is invariant under translations but that rotations spoil the form of the lateral connections unless the orientation tuning of neurons within a hypercolumn is also relabeled (by the amount of rotation). So the Ermentrout-Cowan and the Bressloff-Cowan models both have Euclidean symmetry, but the ways that the Euclidean group acts are different and this leads to different pattern formation results. The end result is that the Bressloff-Cowan model predicts planforms of the type in Figure 11. Note the similarities with the geometric hallucinations reported in Figure 7.



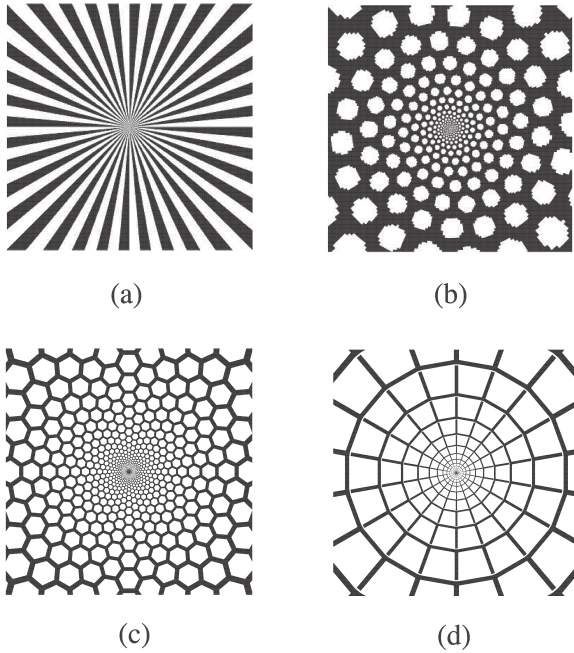
**Figure 10:** Short and long range connections in the visual cortex.  $\cdots$  inhibitory;  
 $-\cdots-$  excitatory.

## 4 Conclusions

We have attempted to show how the existence of symmetry (both in equilibrium and time-periodic states) can help to understand patterns in applications even when the application has no precise mathematical model.

The symmetry description of locomotor central pattern generators leads to a variety of predictions about quadrupedal and bipedal gaits. In this article we described only one: the existence of an unexpected but natural gait – the jump. The proposed structure of CPG models leads to a variety of other predictions (the difference between primary and secondary gaits; the physiological need for each leg to be controlled by two neuron groups; and unexpected properties of centipede primary gaits). See (M. Golubitsky et al., 1999; M. Golubitsky and I. Stewart, 2002).

The symmetry of the primary visual cortex (determined experimentally) led, through symmetry-breaking arguments, to an unexpected correlation between this symmetry and the richness of geometric visual hallucinations. It is important to observe that this correlation can be understood without the need of a detailed model of the cortex V1 – just the symmetry structure that such a model should have.



Visual field planforms

**Figure 11:** Taken from Bressloff et al. 2002

## References

- Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity. *Monkey Striate Cortex* 12, 3139–3161.
- Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., & Wiener, M. C. (2001). Geometric visual hallucinations, Euclidean symmetry, and the functional architecture of striate cortex. *Philosophical Transactions of the Royal Society B* 356, 299–330.
- Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., & Wiener, M. C. (2002). What geometric visual hallucinations tell us about the visual cortex. *Neural Computation* 14, 473–491.
- Bosking, W. H., Zhang, Y., Schofield, B., & Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *The Journal of Neuroscience* 17, 2112–2127.
- Buono, P. L. & Golubitsky, M. (2001). Models of central pattern generators for quadruped locomotion: I. primary gaits. *Journal of Mathematical Biology* 42, 291–326.

- Clottes, J. & Lewis-Williams, D. (1998). *The Shamans of Prehistory: Trance and Magic in the Painted Caves*. New York: Abrams.
- Collins, J. J. & Stewart, I. (1993). Coupled nonlinear oscillators and the symmetries of animal gaits. *Journal of Nonlinear Science* 3. 349–392.
- Collins, J. J. & Stewart, I. (1994). A group-theoretic approach to rings of coupled biological oscillators. *Biological Cybernetics* 71. 95–103.
- Ermentrout, G. B. & Cowan, J. D. (1979). A mathematical theory of visual hallucination patterns. *Biological Cybernetics* 34. 137–150.
- Gambaryan, P. P. (1974). *How Mammals Run: Anatomical Adaptations*. New York: Wiley.
- Golubitsky, M. & Schaeffer, D. G. (1985). *Singularities and Groups in Bifurcation Theory: Vol. 1*. Applied Mathematical Sciences 51, New York: Springer Verlag.
- Golubitsky, M. & Stewart, I. (2002). *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*. Revised Edition. Basel: Birkhäuser.
- Golubitsky, M. & Stewart, I. (2006). Nonlinear dynamics of networks: the groupoid formalism. *Bulletin of the American Mathematical Society* 43. 305–364.
- Golubitsky, M., Stewart, I., Buono, P. L., & Collins, J. J. (1998). A modular network for legged locomotion. *Physica D* 115. 56–72.
- Golubitsky, M., Stewart, I., Buono, P. L., & Collins, J. J. (1999). Symmetry in locomotor central pattern generators and animal gaits. *Nature* 401. 693–695.
- Golubitsky, M., Stewart, I. N., & Schaeffer, D. G. (1988). *Singularities and Groups in Bifurcation Theory: Vol. 2*. Applied Mathematical Sciences 69. New York: Springer Verlag.
- Gorman, M., El-Hamdi, M., & Robbins, K. A. (1994). Experimental Observation of Ordered States in Cellular Flames. *Combustion Science and Technology* 98. 37–45.
- Gorman, M., Hamill, C. F., El-Hamdi, M., & Robbins, K. A. (1994). Rotating and modulated rotating states of cellular flames. *Combustion Science and Technology* 98. 25–35.
- Guckenheimer, J. & Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Applied Mathematical Sciences 42. New York: Springer Verlag.
- Hubel, D. H. & Wiesel, T. N. (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *The Journal of Comparative Neurology* 158. 267–294.
- Hubel, D. H. & Wiesel, T. N. (1974). Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *The Journal of Comparative Neurology* 158. 295–306.
- Hubel, D. H. & Wiesel, T. N. (1974). Ordered arrangement of orientation columns in monkeys lacking visual experience. *The Journal of Comparative Neurology* 158. 307–318.
- Keener, J. & Sneyd, J. (1998). *Mathematical Physiology*. Interdisciplinary Applied Mathematics 8. New York: Springer Verlag.
- Klüver, H. (1966). *Mescal and Mechanisms of Hallucinations*. Chicago: University of Chicago Press.
- Kopell, N. & Ermentrout, G. B. (1988). Coupled oscillators and the design of central pattern generators. *Mathematical Biosciences* 89. 14–23.
- Kopell, N. & Ermentrout, G. B. (1990). Phase transitions and other phenomena in chains of oscillators. *SIAM Journal Applied Mathematics* 50. 1014–1052.
- Michel, L. (1972). Nonlinear group action: Smooth actions of compact Lie groups on manifolds. In: Sen, R. N. & Wiel, C. (eds). *Statistical Mechanics and Field Theory*. Jerusalem: Israel University Press. 133–150.
- Patterson, A. (1992). *Rock Art Symbols of the Greater Southwest*. Boulder: Johnson Books.



- Rand, R. H., Cohen, A. H., & Holmes, P. J. (1988). Systems of coupled oscillators as models of central pattern generators. In: Cohen, A. H., Rossignol, S., & Grillner, S. (eds.). *Neural Control of Rhythmic Movements in Vertebrates*. New York: Wiley. 333–367.
- Sattinger, D. H. (1979). *Group Theoretic Methods in Bifurcation Theory*. Lecture Notes in Mathematics 762. New York: Springer Verlag.
- Schöner, G., Jiang, W. Y., & Kelso, J. A. S. (1990). A synergetic theory of quadrupedal gaits and gait transitions. *Journal of Theoretical Biology* 142. 359–391.
- Siegel, R. K. & Jarvik, M. E. (1975). Drug-Induced Hallucinations in Animals and Man. In: Siegel, R. K. & West, L. J. (eds.). *Hallucinations: Behavior, Experience and Theory*. New York: Wiley. 81–161.
- Thom, R. (1972). *Stabilité structurelle et morphogénèse*. New York: W A Benjamin.
- Zeeman, E. C. (1977). *Catastrophe Theory: Selected papers, 1972–1977*. Oxford: Addison-Wesley.

This work was supported in part by NSF Grant DMS-1008412 to MG and NSF Grant DMS-0931642 to the Mathematical Biosciences Institute.

**Prof. Dr. Martin Golubitsky**

The Ohio State University  
 Mathematical Biosciences Institute  
 Columbus, OH 43215  
 USA  
 mg@mbi.osu.edu



Thomas A. C. Reydon

# Symmetry and the Explanation of Organismal Form

Commentary on Martin Golubitsky

## 1 Introduction

Golubitsky (this volume) presented three examples in which the concepts of symmetry and symmetry breaking, as well as their mathematical formalizations, played an important role in understanding patterns exhibited by physical and biological systems. These examples concerned patterns occurring in burner flames, in animal locomotion and in visual hallucinations. A striking feature of these examples was that the same general mathematical model of symmetry breaking could be applied in all cases, even though the systems under consideration came from quite different realms. Golubitsky's claims were that the mathematics of symmetry and symmetry breaking can help us understand the origins of patterns observed in physical as well as biological systems, and that there is a general "menu of patterns" that encompasses patterns that can be realized in materially very different kinds of systems (this volume; Stewart and Golubitsky, 1993: 186, 207, 218).

The philosophical question that Golubitsky's claims give rise to pertains to mathematical models in general: If there are general mathematical models that apply to materially very different kinds of systems, physical as well as biological ones, and can help us understand how these systems work, then what exactly is the role of such models in understanding and explaining the phenomena under study? What is the epistemic work that such models do in science?

This is a very broad question, which needs to be constrained more. Here, I will only consider one of Golubitsky's examples, namely the explanations of organismal traits such as the various locomotive patterns that animals exhibit. Where do mathematical models of the sort discussed by Golubitsky fit into the larger explanatory structure of biological science? I will begin by addressing the role of mathematical models in biology in general.

## 2 What work do mathematical models do in the biosciences?

Although mathematical models are widespread in biology, the role of mathematics in biology seems quite different from its role in, for example, physics and chemistry. In these latter sciences, mathematical formalisms constitute a core feature of theories and explanations. But this is not so for the principal theories of biology. For example, evolutionary theory and evolutionary explanations, which constitute the backbone of biological science,<sup>1</sup> are often presented in verbal/conceptual form without using much mathematics. Similarly, organismal development is usually explained in terms of the operation of different genes and gene networks without necessarily relying on mathematical formalisms. This is not to say that mathematics is unimportant in developmental and evolutionary biology: it is not (e.g., Rice, 2004), but it does play a less prominent role in biology than in the exact natural sciences. Accordingly, Ernst Mayr (1982: 43) once claimed that progress in biology does not occur by formulating strict laws of the sort found in the physical sciences, but is largely a matter of the articulation and refinement of concepts.

This suggests that mathematical models in biology do not play their main parts in the formulation of explanations. Rather, their main roles might be heuristic. They can aid communication and serve didactical and rhetorical purposes by functioning as metaphors and analogies that represent real systems in ways that are easier to understand than the complex “real thing” (e.g., Stewart, 2003: 184). Moreover, they enable scientists to simulate how systems behave under various conditions in cases in which the “real thing” is difficult to access.

Golubitsky’s example of animal gaits supports this suggestion (this volume; Field and Golubitsky, 1992: 32; Stewart and Golubitsky, 1993: Chapter 8; Golubitsky et al., 1998; 1999; Stewart, 2003; Pinto and Golubitsky, 2006). There, models play two heuristic roles. First, they provide information about how individual animals realize locomotion, thus contributing to the study of how organisms work. According to a widely held (but not uncontroversial – Stewart and Golubitsky, 1993: 201–203) assumption, animal locomotion is controlled by so-called central pattern generators (CPGs), neural networks that control limb motion (Stewart and Golubitsky, 1993: 199–203; Golubitsky et al., 1998: 57; Golubitsky et al., 1999: 693; Stewart, 2003: 197; Pinto and Golubitsky, 2006: 475). CPGs themselves are difficult to study *in vivo* or *in vitro*, so investigators work

---

<sup>1</sup> Allegedly, “nothing in biology makes sense except in the light of evolution” (Dobzhansky, 1964: 449; 1973: 125).

backwards and try to derive information about how CPGs function from observations about the patterns they produce. The models used by Golubitsky and co-workers start from observed symmetries in animal gaits and symmetry breakings that occur in transfers between gaits. From this, the possible structures of the underlying CPGs are inferred, guided by the thought that the observed symmetries and symmetry breakings must correspond to those that an abstract network of a limited number of nodes can produce. The observed symmetries thus allow inferences about the symmetries of the underlying networks: “symmetry can be used to infer a plausible class of CPG network architectures from observed patterns of animal gaits” (Golubitsky et al., 1999: 693). In turn, from the symmetries of these general network architectures possible gaits can be predicted and looked for in animals in nature.<sup>2</sup>

Second, the relations between the various models of animal gaits can be used as indirect evidence for possible evolutionary scenarios (Pinto and Golubitsky, 2006: 487; Stewart, 2003: 196). The number of steps required to get from one set of gaits to another can be interpreted as an indication of the number of steps that evolution must have taken on its way from a taxon exhibiting one set of gaits to a taxon exhibiting the other set. For example, the steps needed to get from the set of gaits characteristic of quadrupedal locomotion to the set for bipedal locomotion can be taken to indicate the steps taken in the evolution of bipedal organisms from quadrupeds. Thus, mathematical models can provide clues about the evolutionary distance between and evolutionary history of taxa.

In both these cases, the inference is toward a class of possibilities (a class of possible CPG structures and a class of possible evolutionary routes). The models provide clues about which architectures or routes are possible, but not about the *actual* architectures or routes involved and thus don’t provide any concrete explanatory details. The question thus remains open whether mathematical models can be more than heuristic tools and might perform “proper” explanatory roles. I will address this question by considering the search for a theory of organismal form.

---

<sup>2</sup> Golubitsky’s example of visual hallucinations works in the same manner (Bressloff et al., 2001: 323–326; Bressloff et al., 2002: 476–477). The question is which neural network architectures are required to produce the variety of geometrical patterns found in visual hallucinations. This is answered by relating the observed symmetries of hallucination patterns to the symmetries that a producing network must possess. In this way the *possible* architectures of the visual cortex area responsible for producing visual hallucinations are inferred from the *actual* patterns of observed hallucinations.

### 3 What natural selection does not explain

Among the principal questions of biological science are why we have the organismal diversity that we do (rather than a different diversity) and why the organisms we find around us have the traits they do, instead of other possible traits they might have exhibited (and that sometimes organisms of different species *do* exhibit). Ever since Darwin's work an important part of the answers to these questions is given in terms of natural selection. But it has long been clear that selection constitutes only part of the answer.

In the first place, not *all* organismal traits are necessarily explained by selection, as paleontologist Stephen Jay Gould and geneticist Richard Lewontin pointed out in their famous “spandrels” paper (Gould and Lewontin, 1979). They criticized a procedure commonly followed by biologists, namely to break organisms down into discrete traits and to propose a separate adaptive story for each trait. Each trait's presence is then explained as a consequence of some function that it performed in ancestral organisms, endowing these with a selective advantage over organisms not possessing the trait in question. The underlying assumption is that “natural selection [is] so powerful and the constraints upon it so few that direct production of adaptation through its operation becomes the primary cause of nearly all organic form, function, and behaviour” (Gould and Lewontin, 1979: 584–585). However, Gould and Lewontin argued, this assumption stands unsupported: many organismal traits might be correctly explained as products of evolution by means of natural selection, but not necessarily all or nearly all traits are. Other explanatory factors besides natural selection, such as constraints on organismal development, also play important roles and may outweigh the explanatory importance of selection. Thus, Gould and Lewontin argued in favor of a pluralistic approach to biological explanation in which a plurality of explanatory factors can be invoked when explaining biodiversity and organismal traits. As they pointed out (Gould and Lewontin 1979: 589), this is in line with Darwin's own view “that Natural Selection has been the main but not exclusive means of modification” (Darwin, 1859: 6).

Furthermore, even for traits that are correctly explained as products of natural selection, selection is only part of the answer. Selection explains the trait's *presence* and its adaptive aspects, but there is more to say. Soon after the publication of the *Origin of Species*, biologists have begun to criticize Darwin's theory for addressing the spread and persistence of traits through ancestor-descendant lineages but not being able to explain how these traits arise in the first place (see Reydon, 2011). The criticism, which is also voiced by some contemporary biologists (Fontana and Buss, 1994; Gilbert, 2000), is that even if natural selection can

cause the differential reproduction of organism types with varying traits, it needs material to work with: natural selection filters, but it does not create new traits.

These two criticisms constitute the motivation behind a tradition of work in biology aiming to develop a theory of the origins of organismal forms, where ‘form’ is understood broadly as encompassing the shapes of organisms as well as their other physical and behavioral traits. The theory sought after should explain the origins of organismal traits and complement the theory of selection, which explains their preservation and spread.

## 4 Growth and form: D’Arcy Thompson’s project

A key figure in the quest for a theory of organismal form was zoologist D’Arcy Wentworth Thompson. In his *On Growth and Form*, Thompson developed the project of comparing organismal forms to forms and patterns found in non-living systems and understanding these as instances of the same phenomena. The central thought in Thompson’s book is that the principal causes of organismal forms are physical forces, such that organismal traits should be explained by taking recourse to general physical and chemical principles rather than selection and adaptation. Thompson thought of natural selection as a mere filter that could not create evolutionary novelty and thus could not explain organismal form (Bonner, 1992: xvii).

In a famous example, he compared the shapes of jellyfish to the shapes that liquid drops assume when falling through other liquids and suggested that both phenomena might be susceptible to the same explanation (Thompson, 1942: 392–398). Jellyfish here are modeled as expanding drops of a fluid with a different density than the water in which they are immersed and the observed shapes are explained as consequences of the operation of the physical laws that govern the flow of fluids in fluids.<sup>3</sup> It is unclear, however, exactly how much explanatory work Thompson’s mathematical models do. For instance, Thompson writes:

[W]e may use a hanging drop, which, while it sinks, remains suspended to the surface ... [T]he figure so produced, in either case, is closely analogous to that of a medusa or jellyfish ... *It is hard to say how much or little all these analogies imply.* But they indicate, at the very least, how certain simple organic forms *might be naturally assumed by one fluid mass within another*, when gravity, surface tension and fluid friction play their part (Thompson, 1942: 395–398; emphasis added).

---

<sup>3</sup> Note that another of Thompson’s (1942: 39–50) examples concerned animal locomotion and flight.

Although Thompson was careful not to imply too much, this quotation does suggest that he took the analogy as having *some* explanatory value in that the various shapes of jellyfish can be explained as what is bound to occur for particular fluids under particular conditions.

Similarly to Golubitsky's models, Thompson's models take recourse to physical laws to map out the spectrum of what is possible under various conditions (Bonner, 1992: xxii). In this respect, the laws of physics function in the same way in explanations of organismal form as in explanations of phenomena in the non-living realm: in both cases there are general physical principles that apply universally and determine what is bound to occur in such-and-such kinds of systems under such-and-such conditions, irrespective of the systems' material bases. As Thompson writes at the end of his book: "So the living and the dead, things animate and inanimate ... are bound alike by physical and mathematical law" (Thompson, 1942: 1097).

This motif is found elsewhere too. For example, zoologist Rupert Riedl remarked that "[t]he living world happens to be crowded by universal patterns of organization which ... find no direct explanation through environmental conditions or adaptive radiation, but exist primarily through universal requirements which can only be expected under the systems conditions of complex organization itself" (Riedl, in Gould and Lewontin, 1979: 594). In a similar spirit, mathematician (and frequent collaborator of Golubitsky's) Ian Stewart remarked about the observed symmetry breakings in the developmental cycle of the alga *Acetabularia acetabulum* that these are the same as found in a particular type of fluid flow, "*as they should be* since such patterns are universal in cylindrically symmetric systems" (Stewart, 2003: 190; emphasis added). And it seems to me to be the motif underlying Golubitsky's suggestion that there is a general "menu of patterns" that can be realized in materially very different kinds of systems found in different realms in nature (this volume; Stewart and Golubitsky, 1993: 186, 207, 218).

Invoking such universal patterns that can be captured in mathematical models of symmetry and symmetry breakings does not explicate what is *actually* the case in a system under study, as it abstracts away from the system's characteristics. It narrows down the set of possible explanations of the phenomenon under study to a limited number of possible scenarios. On some accounts of explanation this could be accepted as "proper" scientific explanation and Thompson's and Golubitsky's models would count as "how possibly" explanations (O'Hara, 1988; Brandon, 1990; Resnik, 1991; Reiner, 1993). However, whether "how possibly" explanations should be accepted as "proper" scientific explanations is still a controversial issue in the philosophy of science.

## 5 Conclusion

As Golubitsky showed, symmetry breaking is common in the living world, e.g., in animal locomotion or organismal growth.<sup>4</sup> In Thompson's project, too, the concept of symmetry played an important role: "In all cases where the principle of maxima and minima comes into play [...] the configurations so produced are characterized by obvious and remarkable *symmetry*. Such symmetry is highly characteristic of organic forms and is rarely absent in living things" (Thompson, 1942: 357). If this is right, there clearly must be epistemic work to do for the concepts of symmetry and symmetry breaking and their mathematical formalizations in explanations of organismal form. But there are good reasons to think of this work as not being explanatory in and by itself.

Even though mathematical models of symmetry and symmetry breaking seem to provide "how possibly" explanations, the mathematics *itself* does not provide explanatory force: the applicable physical laws and system specifications do (cf. Stewart, 2003: 191). Similarly, symmetry breaking itself does not explain much. The explanatory work is done by the causes *underlying* symmetry breakings, i.e., the physical laws that govern particular kinds of systems and the slight imbalances in an overall symmetrical system that at some point causes the breaking of its symmetry (Stewart, 2003: 188). That the same mathematical model applies to a number of very different systems merely indicates that in all these systems the same physical laws are involved. Mathematical models of symmetries and symmetry breakings do not capture the complexity of the systems under study, but abstract away from much detail, allowing us to focus on the relevant overall patterns and to identify the relevant underlying laws. While this is important to gain insight into what could occur in the system under consideration, actual explanations of concrete phenomena will need to specify the details of the system itself.

Golubitsky's examples showed how models of symmetries and symmetry breakings provide clues about what might possibly be the case in the systems under study. The models describe how organismal function, development and evolution are constrained by the general laws of physics and chemistry, making some traits possible and others impossible (cf. Stewart, 2003: 200). One might interpret such models as adding "how possibly" explanations to the "how and why actually" explanations of functional, developmental and evolutionary biology. But in my view their role in fact is more heuristic in nature and it is to be doubted whether such "how possibly" explanations should count as "proper" scientific explanations on an equal level with other explanations in biology.

---

<sup>4</sup> Another example: non-spherically-symmetrical starfish develop from spherically symmetrical eggs (Field and Golubitsky, 1992: 32).

## References

- Bonner, J. T. (1992). The editor's introduction. In: Thompson, D'A. W.: *On Growth and Form, An Abridged Edition Edited by John Tyler Bonner*, Cambridge: Cambridge University Press. xiv–xxii.
- Brandon, R. N. (1990). *Adaptation and Environment*. Princeton (NJ): Princeton University Press.
- Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., & Wiener, M. C. (2001). Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. *Philosophical Transactions of the Royal Society of London B* 356. 299–330.
- Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., & Wiener, M. C. (2002). What geometric visual hallucinations tell us about the visual cortex. *Neural Computation* 14. 473–491.
- Darwin, C. R. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.
- Dobzhansky, T. (1964). Biology, molecular and organismic. *American Zoologist* 4. 443–452.
- Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher* 35. 125–129.
- Field, M. & Golubitsky, M. (1992). *Symmetry in Chaos: A Search for Pattern in Mathematics, Art and Nature*. Oxford: Oxford University Press.
- Fontana, W. & Buss, L. W. (1994). “The arrival of the fittest”: Toward a theory of biological organization. *Bulletin of Mathematical Biology* 56. 1–64.
- Gilbert, S. F. (2000). Genes classical and genes developmental. In: Beurton, P., Falk, R., & Rheinberger, H.-J. (eds.): *The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives*. Cambridge: Cambridge University Press. 178–192.
- Golubitsky, M., Stewart, I., Buono, P.-L., & Collins, J. J. (1998). A modular network for legged locomotion. *Physica D* 115. 56–72.
- Golubitsky, M., Stewart, I., Buono, P.-L., & Collins, J. J. (1999). Symmetry in locomotor central pattern generators and animal gaits. *Nature* 401. 693–695.
- Gould, S. J. & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London B* 205. 581–598.
- Mayr, E. (1982). *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Cambridge (MA): Harvard University Press.
- Pinto, C. M. A. & Golubitsky, M. (2006). Central pattern generators for bipedal locomotion. *Journal of Mathematical Biology* 53. 474–489.
- O'Hara, R. J. (1988). Homage to Clio, or, toward an historical philosophy for evolutionary biology. *Systematic Biology* 37. 142–155.
- Reiner, R. (1993). Necessary conditions and explaining how-possibly. *Philosophical Quarterly* 43. 58–69.
- Resnik, D. B. (1991). How-possibly explanations in biology. *Acta Biotheoretica* 39. 141–149.
- Reydon, T. A. C. (2011). The arrival of the fittest *what?* In: Dieks, D., Gonzalez, W. J., Hartmann, S., Uebel, T., & Weber, M. (eds.): *Explanation, Prediction, and Confirmation*. Dordrecht: Springer. 223–237.
- Rice, S. H. (2004). *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sunderland (MA): Sinauer.



- Stewart, I. (2003). Broken symmetries and biological patterns. In: Kumar, S. & Bentley, P. J. (eds.): *On Growth, Form and Computers*. London and San Diego: Elsevier Academic Press. 181–202.
- Stewart, I. & Golubitsky, M. (1993). *Fearful Symmetry: Is God a Geometer?* Harmondsworth: Penguin.
- Thompson, D'A. W. (1942). *On Growth and Form: New Edition*. Cambridge: Cambridge University Press.

**Prof. Dr. Thomas A. C. Reydon**

Leibniz University Hannover  
Institute of Philosophy  
Im Moore 21  
30167 Hannover  
Germany  
[reydon@ww.uni-hannover.de](mailto:reydon@ww.uni-hannover.de)



Dirk Helbing

# Pluralistic Modeling of Complex Systems

## 1 Introduction

When the father of sociology, August Comte, came up with the idea of a “social physics”, he hoped that the puzzles of social systems could be revealed with a natural science approach (Comte, A., 1856). However, progress along these lines was very difficult and slow. Today, most sociologists do not believe in his positivistic approach anymore. The question is whether this proves the failure of the positivistic approach or whether it just shows that social scientists did not use the right methods so far. After all, social scientists rarely have a background in the natural sciences, while the positivistic approach has been most successful in fields like physics, chemistry, or biology.

In fact, recently, new scientific communities are developing, and they are growing quickly. They call themselves socio-physicists, mathematical sociologists, computational social scientists, agent-based modelers, complexity or network scientists. Researchers from the social sciences, physics, computer science, biology, mathematics, and artificial intelligence research are addressing the challenges of social and economic systems with mathematical or computational models and lab or web experiments. Will they end up with resignation in view of the complexity of social and economic systems, or will they manage to push our knowledge of social systems considerably beyond what was imaginable even a decade ago? Will August Comte’s vision of sociology as “the queen of the sciences” (Comte, A., 1830–1842) finally become true?

My own judgement is that it is less hopeless to develop mathematical models for social systems than most social scientists usually think, but more difficult than most natural scientists imagine. The crucial question is, how one can make substantial progress in a field as complicated and multi-faceted as the social sciences, and how the current obstacles can be overcome. And what are these obstacles, after all? The current contribution tries to make the controversial issues better understandable to scientific communities with different approaches and backgrounds. While each of the points may be well-known to some scientists, they are probably not so obvious for others. Putting it differently, this contribution tries to build bridges between different disciplines interested in similar subjects, and to make thoughts understandable to scientific communities with different points of views.

A dialogue between social, natural and economic sciences seems to be desirable not only for the sake of an intellectual exchange on fundamental scientific

problems. It also appears that science is behind the pace of upcoming socio-economic problems, and that we need to become more efficient in addressing practical problems (Helbing, D., 2010, *Grand socio-economic challenges*). President Lee C. Bollinger of New York's prestigious Columbia University formulated the challenge as follows:

The forces affecting societies around the world (...) are powerful and novel. The spread of global market systems (...) are (...) reshaping our world (...), raising profound questions. These questions call for the kinds of analyses and understandings that academic institutions are uniquely capable of providing. Too many policy failures are fundamentally failures of knowledge.<sup>1</sup>

The fundamental and practical scientific challenges require from us to do everything we can to find solutions, and not to give up before the limits or failure of a scientific approach have become obvious. As will be argued in Sec. 5, different methods should be seen complementary to each other and, even when inconsistent, may allow one to get a better picture than any single method can do, no matter how powerful it may seem.

## 2 Particular Difficulties of Modeling Socio-Economic Systems

When speaking about socio-economic systems in the following, it could be anything from families over social groups or companies up to countries, markets, or the world economy including the financial system and the labor market. The constituting system elements or system components would be individuals, groups, or companies, for example, depending on the system under consideration and the level of description one is interested in.

On the macroscopic (systemic) level, social and economic systems have some features that seem to be similar to properties of some physical or biological systems. One example is the hierarchical organization. In social systems, individuals form groups, which establish organizations, companies, parties, etc., which make up states, and these build communities of states (like the United States or the European Union, for example). In physics, elementary particles form atoms, which create molecules, which may form solid bodies, fluids or gases, which

---

<sup>1</sup> L. C. Bollinger, announcing the Columbia committee on global thought, see <http://www.columbia.edu/cu/president/docs/communications/2005-2006/051214-committee-global-thought.html>.

together make up our planet, which belongs to a solar system, and a galaxy. In biology, cells are composed of organelles, they form tissues and organs, which are the constituting parts of living creatures, and these make up ecosystems.

Such analogies are certainly interesting and have been discussed, for example, by Herbert Spencer (Spencer, H., 1898) and later on in systems theory (Bertalanffy, L. von, 1968). It is not so obvious, however, how much one can learn from them. While physical systems are often well understood by mathematical models, biological and socio-economic systems are usually not. This often inspires physicists to transfer their models to biological and socioeconomic models (see the discussion in Sec. 4.4), while biologists, social scientists, and economists often find such attempts “physicalistic” and inadequate. In fact, social and economic systems possess a number of properties, which distinguish them from most physical ones:

1. the number of variables involved is typically (much) larger (considering that each human brain contains about 1000 billion neurons),
2. the relevant variables and parameters are often unknown and hard to measure (the existence of “unknown unknowns” is typical),
3. the time scales on which the variables evolve are often not well separated from each other,
4. the statistical variation of measurements is considerable and masks laws of social behavior, where they exist (if they exist at all),
5. frequently there is no ensemble of equivalent systems, but just one realization (one human history),
6. empirical studies are limited by technical, financial, and ethical issues,
7. it is difficult or impossible to subdivide the system into simple, non-interacting subsystems that can be separately studied,
8. the observer participates in the system and modifies social reality,
9. the non-linear and/or network dependence of many variables leads to complex dynamics and structures, and sometimes paradoxical effects,
10. interaction effects are often strong, and emergent phenomena are ubiquitous (hence, not understandable by the measurement and quantification of the individual system elements),
11. factors such as a large degree of randomness and heterogeneity, memory, anticipation, decision-making, communication, consciousness, and the relevance of intentions and individual interpretations complicate the analysis and modeling a lot,
12. the same applies to human features such as emotions, creativity, and innovation,

13. the impact of information is often more decisive for the behavior of a socio-economic system than physical aspects (energy, matter) or our biological heritage,
14. the “rules of the game” and the interactions in a social or economic system may change over time, in contrast to what we believe to be true for the fundamental laws and forces of physics,
15. in particular, social systems are influenced by normative and moral issues, which are variable.

For such reasons, social systems are the most complex systems we know. They are certainly more complex than physical systems are. As a consequence, a considerable fraction of sociologists thinks that mathematical models for social systems are destined to fail, while most economists and many quantitatively oriented social scientists seem to believe in models with many variables. Both is in sharp contrast to the often simple models containing a few variables only, which physicists tend to propose. So, who is right? The following discussion suggests that this is the wrong question. We will therefore discuss why different scientists, who apparently deal with the same research subject, come to so dramatically different conclusions.

It is clear that this situation has some undesirable side effects: Scientists belonging to different schools of thought often do not talk to each other, do not learn from each other, and probably reject each others' project proposals more frequently. It is, therefore, important to make the approach of each school understandable to the others.

## 3 Modeling Philosophies

### 3.1 Qualitative Descriptions

Many social scientists think that the fifteen challenges listed above are so serious that it is hopeless to come up with mathematical models for social systems. The basic philosophy seems to be that all models are wrong. Thus, a widespread approach is to work out narratives, i.e. to give a qualitative (non-mathematical and non-algorithmic) description of reality that is as detailed as possible. This may be compared with a naturalist painting.

Narratives are important, as they collect empirical evidence and create knowledge that is essential for modelers sooner or later. Good models require several steps of intellectual digestion, and the first and very essential one is to

create a picture of the system one is interested in and to make sense of what is going on in it. This step is clearly indispensable. Nevertheless, the approach is sometimes criticized for reasons such as the following:

- Observation, description, and interpretation are difficult to separate from each other, since they are typically performed by the same brain (of a single scientist). Since these processes strongly involve the observer, it is hard or even impossible to provide an objective description of a system at this level of detail. Therefore, different scientists may analyze and interpret the system in different, subjective ways. What is an important aspect for one observer may be an irrelevant detail for another, or may even be overlooked. There is a saying that “one misses the forest for the trees”, i.e. details may hide the bigger picture or the underlying mechanisms. In the natural sciences, this problem has been partially overcome by splitting up observation, description, and interpretation into separate processes: measurements, statistical analysis, and modeling attempts. Many of these steps are supported by technical instruments, computers, and software tools to reduce the individual element and subjective influence. Obviously, this method cannot be easily transferred to the study of social systems, as individuals and subjective interpretations can have important impacts on the overall system.
- Despite its level of detail, a narrative is often not suited to be translated into a computer program that would reproduce the phenomena depicted by it. When scientists try to do so, in many cases it turns out that the descriptions are ambiguous, i.e. not detailed enough to come up with a unique computer model. In other words, different programmers would end up with different computer models, producing different results. Therefore, Joshua Epstein claims: “If you didn’t grow it, you didn’t explain it” (Epstein, J. M., 2006) (where “grow” stands here for “simulate in the computer”). For example, if system elements interact in a non-linear way, i.e. effects are not proportional to causes, there are many different possibilities to specify the non-linearity: is it a parabola, an exponential dependence, a square root, a logarithm, a power law, ...? Or when a system shows partially random behavior, is it best described by additive or multiplicative noise, internal or external noise? Is it chaotic or turbulent behavior, or are the system elements just heterogeneous? It could even be a combination of several options. What differences would these various possibilities make?

## 3.2 Detailed Models

In certain fields of computational social science or economics, it is common to develop computer models that grasp as many details as possible. They would try to implement all the aspects of the system under consideration, which are known to exist. In the ideal case, these facts would be properties, which have been repeatedly observed in several independent studies of the kind of system under consideration, preferably in different areas of the world. In some sense, they would correspond to the overlapping part of many narratives. Thus, one could assume that these properties would be characteristic features of the kind of system under consideration, not just properties of a single and potentially quite particular system.

Although it sounds logical to proceed in this way, there are several criticisms of this approach:

- In case of many variables, it is difficult to specify their interdependencies in the right way. (Just remember the many different possibilities to specify non-linear interactions and randomness in the system.)
- Some models containing many variables may have a large variety of different solutions, which may be highly dependent on the initial or boundary conditions, or the history of the system. This particularly applies to models containing non-linear interactions, which may have multiple stable solutions or non-stationary ones (such as periodic or non-periodic oscillations), or they may even show chaotic behavior. Therefore, depending on the parameter choice and the initial condition, such a model could show virtually any kind of behavior. While one may think that such a model would be a flexible world model, it would in fact be just a fit model. Moreover, it would probably not be very helpful to understand the mechanisms underlying the behavior of the system. As some people say: “A model containing more than 3 parameters can fit an elephant” (Dyson, F., 2004), which wants to express that a model with many parameters can fit anything and explains nothing. This is certainly an extreme standpoint, but there is some truth in it.
- When many variables are considered, it is hard to judge which ones are independent of each other and which ones are not. If variables are mutually dependent, one effect may easily be considered twice in the model, which would lead to biased results. Dependencies among variables may also imply serious problems in the process of parameter calibration. The problem is known, for example, from sets of linear equations containing collinear variables.
- Models with many variables, particularly non-linear ones, may be sensitive to the exact specification of parameters, initial, or boundary conditions, or to



small random effects. Phenomena like hysteresis (history-dependence) (Mayergoyz, I. D., 2003), phase transitions (Stanley, H. E., 1987) or “catastrophes” (Zeeman, E. C. (ed.), 1977), chaos (Schuster, H. G., and Just, W., 2005), or noise-induced transitions (Horsthemke, W., and Lefever, R., 1983) illustrate this clearly.

- The parameters, initial and boundary conditions of models with many variables are hard to calibrate. If small (or no) data sets are available, the model is under-specified, and the remaining data must be estimated based on “expert knowledge”, intuition or rules of thumb, but due to the sensitivity problem, the results may be quite misleading. The simulation of many scenarios with varying parameters can overcome the problem in part, as it gives an idea of the possible variability of systemic behaviors. However, the resulting variability can be quite large. Moreover, a full exploration of the parameter space is usually not possible when a model contains many parameters, not even with supercomputers.
- In models with many variables, it is often difficult to identify the mechanism underlying a certain phenomenon or system behavior. The majority of variables may be irrelevant for it. However, in order to understand a phenomenon, it is essential to identify the variables and interactions (i.e. the interdependencies among them) that matter.

### 3.3 Simple Models

Simple models try to avoid (some of) the problems of detailed models by restricting themselves to a minimum number of variables needed to reproduce a certain effect, phenomenon or system behavior. They are aiming at a better understanding of so-called “stylized facts”, i.e. simplified, abstracted, or “idealtypical” observations (the “essence”). For example, while detailed descriptions pay a lot of attention to the particular content of social norms or opinions and how they change over time in relation to the respective cultural setting, simple models abstract from the content of social norms and opinions. They try to formulate general rules of how social norms come about or how opinions change, independently of their content, with the aim of understanding why these processes are history-dependent (“hysteretic”) and in what way they depend on microscopic and macroscopic influences.

It is clear that simple models do not describe (and do not even want to describe) all details of a system under consideration, and for this reason they are also called minimal or toy models sometimes. The philosophy of this approach may be represented by a few quotes. The “KISS principle” of model build-

ing demands to “keep it simple and straightforward”<sup>2</sup>. This is also known as Occam’s (or Ockham’s) razor, or as principle of parsimony. Albert Einstein as well demanded (Einstein, A., 1934): “Make everything as simple as possible, but not simpler”.

A clear advantage of simple models is that they may facilitate an analytical treatment and, thereby, a better understanding. Moreover, it is easy to extend simple models in a way that allows one to consider heterogeneity among the system components. This supports the consideration of effects of individuality and the creation of simple “ecological models” for socio-economic systems. Nevertheless, as George Box puts it: “Essentially, all models are wrong, but some are useful” (Box, G. E. P., and Draper, N. R., 1987).

The last quote touches an important point. The choice of the model and its degree of detail should depend on the purpose of a model, i.e. its range of application. For example, there is a large variety of models used for the modeling and simulation of freeway traffic. The most prominent model classes are “microscopic” car-following models, focussing on the interaction of single vehicles, “mesoscopic” gas-kinetic models, describing the change of the velocity distribution of vehicles in space and time, “macroscopic” fluid-dynamic models, restricting themselves to changes of the average speed and density of vehicles, and cellular automata, which simplify microscopic ones in favor of simulation speed. Each type of model has certain ranges of application. Macroscopic and cellular automata models, for example, are used for large-scale traffic simulations to determine the traffic situation on freeways and perform short-term forecasts, while microscopic ones are used to study the interaction of vehicles and to develop driver assistance systems. For some of these models, it is also known how they are mathematically connected with each other, i.e. macroscopic ones can be derived from microscopic ones by certain kinds of simplifications (approximations) (Helbing, D., 2009, *Derivation*; Helbing, D., 2001).

The main purpose of models is to guide people’s thoughts. Therefore, models may be compared with city maps. It is clear that maps simplify facts; otherwise they would be quite confusing. We do not want to see any single detail (e.g. each tree) in them. Rather we expect a map to show the facts we are interested in, and depending on the respective purpose, there are quite different maps (showing streets, points of interest, topography, supply networks, industrial production, mining of natural resources, etc.).

One common purpose of models is prediction, which is mostly (mis)understood as “forecast”, while it often means “the identification of implications

---

<sup>2</sup> “KISS principle” at Wikipedia.org. See [http://en.wikipedia.org/wiki/KISS\\_principle](http://en.wikipedia.org/wiki/KISS_principle).

regarding how a system is expected to behave under certain conditions". It is clear that, in contrast to the motion of a planet around the sun, the behavior of an individual can hardly be forecasted. Nevertheless, there are certain tendencies or probabilities of doing certain things, and we usually have our hypotheses of what our friends, colleagues, or family members would do in certain situations. It turns out that, when many people interact, the aggregate behavior can sometimes be quite predictable. For example, the "wisdom of crowds" is based on the statistical law of large numbers (Galton, F., 1907), according to which individual variations (here: the independent estimation of facts) are averaged out.

Furthermore, interactions between many individuals tend to restrict the degree of freedom regarding what each individual can or will do. This is, why the concept of "social norms" is so important. Another example is the behavior of a driver, which is constrained by other surrounding vehicles. Therefore, the dynamics of traffic flows can be mathematically well understood (Helbing, D., 2001).<sup>3</sup> Nevertheless, one cannot exactly forecast the moment in which free traffic flow breaks down and congestion sets in, and therefore, one cannot forecast travel times well. The reason for this is the history-dependent dynamics, which makes it dependent on random effects, namely on the size of perturbations in the traffic flow. However, what can be predicted is what are the possible traffic states and what are conditions under which they can occur. One can also identify the probability of traffic flows to break down under certain flow conditions, and it is possible to estimate travel times under free and congested flow conditions, given a measurement of the inflows. The detail that cannot be forecasted is the exact moment in which the regime shift from free to congested traffic flow occurs, but this detail has a dramatic influence on the system. It can determine whether the travel time for a certain freeway section is 2 minutes or 20 minutes.

However, it is important to underline that, in contrast to what is frequently stated, the purpose of developing models is not only prediction. Joshua Epstein, for example, discusses 16 other reasons to build models, including explanation, guiding data collection, revealing dynamical analogies, discovering new questions, illuminating core uncertainties, demonstrating tradeoffs, training practitioners, and decision support, particularly in crises (Epstein, J. M., 2008).

Of course, not everybody favors simple models, and typical criticisms of them are:

- It is usually easy to find empirical evidence, which is not compatible with simple models (even though, to be fair, one would have to consider the purpose they have been created for, when judging them). Therefore, one can

---

<sup>3</sup> Helbing, D. et al. See collection of publications on analytical traffic flow theory at <http://www.soms.ethz.ch/research/traffictheory>.

say that simple models tend to over-simplify things and leave out more or less important facts. For this reason, they may be considered inadequate to describe a system under consideration.

- Due to their simplicity, it may be dangerous to take decisions based on their implications.
- It may be difficult to decide, what the few relevant variables and parameters are, which a simple model should consider. Scientists may even disagree about the stylized facts to model.
- Simple models tend to reproduce a few stylized facts only and are often not able to consistently reproduce a large number of observations. The bigger picture and the systemic view may get lost.
- Making simple models compatible with a long list of stylized facts often requires to improve or extend the models by additional terms or parameter dependencies. Eventually, this improvement process ends up with detailed models, leaving one with the problems specified there (see Sec. 3.2).
- Certain properties and behaviors of socio-economic systems may not be understandable with methods, which have been successful in physics: Subdividing the system into subsystems, analyzing and modeling these subsystems, and putting the models together may not lead to a good description of the overall system. For example, several effects may act in parallel and have non-separable orders of magnitude. This makes it difficult or impossible to start with a zeroth or first order approximation and to improve it by adding correction terms (as it is done, for example, when the falling of a body is described by the effect of gravitational acceleration plus the effect of air resistance). Summing up the mathematical terms that describe the different effects may not converge. It is also not clear whether complex systems can be always understood via simple principles, as the success of complexity science might suggest. Some complex systems may require complex models to explain them, and there may even be phenomena, whose complexity is irreducible. Turbulence (Davidson, P. A., 2004) could be such an example. While it is a long-standing problem that has been addressed by many bright people, it has still not been explained completely.

It should be added, however, that we do not know today, whether the last point is relevant, how relevant it is, and where. So far, it is a potential problem one should be aware of. It basically limits the realm, in which classical modeling will be successful, but we have certainly not reached these limits, yet.

### 3.4 Modeling Complex Systems

Modeling socio-economic systems is less hopeless than many social scientists may think (Weidlich, W., 2006). In recent years, considerable progress has been made in a variety of relevant fields, including

- experimental research (Kagel, J. H., and Roth, A. E., 1995; Guala, F., 2005; Helbing, D., and Yu, W., 2010),
- data mining (Maimon, O., and Rokach, L., 2005),
- network analysis (Jackson, M. O., 2008),
- agent-based modeling (Epstein, J. M., 2006; Gilbert, N. (ed.), 2010),
- the theory of complex systems (including self-organization phenomena and chaos) (Miller, J. H. and Page, S. E., 2007),
- the theory of phase transitions (Stanley, H. E., 1987) (“catastrophes” (Zeeman, E. C. (ed.), 1977)), critical phenomena (Sornette, D., 2006), and extreme events (Albeverio, S., et al. (eds.), 2005), and
- the engineering of intelligent systems (Floreano, D., and Matussi, C., 2005; Nolfi, S., and Floreano, D., 2000).

These fields have considerably advanced our understanding of complex systems. In this connection, one should be aware that the term “complexity” is used in many different ways. In the following, we will distinguish three kinds of complexity:

1. structural,
2. dynamical, and
3. functional complexity.

One could also add algorithmic complexity, which is given by the amount of computational time needed to solve certain problems. Some optimization problems, such as the optimization of logistic or traffic signal operations, are algorithmically complex (Helbing, D., et al., 2009).

Linear models are not considered to be complex, no matter how many terms they contain. An example for structural complexity is a car or airplane. They are constructed in a way that is dynamically more or less deterministic and well controllable, i.e. dynamically simple, and they also serve relatively simple functions (the motion from a location A to another location B). While the acceleration of a car or a periodic oscillation would be an example for a simple dynamics, examples for complex dynamical behavior are non-periodic changes, deterministic chaos, or history-dependent behaviors. Complex dynamics can already be produced by simple sets of non-linearly coupled equations. While a planet orbiting around the sun follows a simple dynamics, the interaction of three celestial

bodies can already show a chaotic dynamics. Ecosystems, the human body or the brain would be functionally complex systems. The same would hold for the world wide web, financial markets, or running a country or multi-national company.

While the interrelation between function, form and dynamics still poses great scientific challenges, the understanding of structurally or dynamically complex systems has significantly progressed. Simple agent-based models of systems with a large number of interacting system elements (be it particles, cars, pedestrians, individuals, or companies) show properties, which remind of socio-economic systems. Assuming that these elements mutually adapt to each other through non-linear or network interactions (i.e. that the elements are influenced by their environment while modifying it themselves), one can find a rich, history-dependent system behavior, which is often counter-intuitive, hardly predictable, and seemingly uncontrollable. These models challenge our common way of thinking and help to grasp behaviors of complex systems, which are currently a nightmare for decision-makers.

For example, complex systems are often unresponsive to control attempts, while close to “critical points” (also known as “tipping points”), they may cause sudden (and often unexpected) phase transition (so-called “regime shifts”). These correspond to discontinuous changes in the system behavior. The breakdown of free traffic flow would be a harmless example, while a systemic crisis (such as a financial collapse or revolution) would be a more dramatic one. Such systemic crises are often based on cascade spreading through network interactions (Helbing, D., 2009, *System risks*). Complex adaptive systems also allow one to understand extreme events as a result of strong interactions in a system (rather than as externally caused shocks). Furthermore, the interaction of many system elements may give rise to interesting self-organization phenomena and emergent properties, which cannot be understood from the behaviors of the single elements or by adding them up. Typical examples are collective patterns of motion in pedestrian crowds or what is sometimes called “swarm intelligence” (Mousaid, M., et al., 2009).

Considering this, it is conceivable that many of today’s puzzles in the social sciences may one day be explained by simple models, namely as emergent phenomena resulting from interactions of many individuals and/or other system elements. Note that emergent phenomena cannot be explained by linear models (which are most common in many areas of quantitative empirical research in the social sciences and economics). Unfortunately, there is no standard way to set up models of emergent phenomena. On the one hand, there are many possible kinds of non-linear functional dependencies (“interactions”) (see the end of Sec. 3.1). On the other hand, model assumptions that appear plausible do often not produce the desired or expected effects.

In spite of these difficulties, taking time-dependent change into account, a non-linear coupling of variables, spatial or network interactions, randomness, and/or correlations (i.e. features that many social and economic models currently do not consider to the necessary extent), can sometimes deliver unexpected solutions of long-standing puzzles. For example, it turns out that representative agent models (which are common in economics) can be quite misleading, as the same kinds of interactions among the system components can imply completely different or even opposite conclusions, when interactions take place in a socio-economic network rather than with average (or randomly chosen) interaction partners (Helbing, D., et al., 2010). In other words, models often produce counter-intuitive results, when spatio-temporal or network interactions are relevant. Therefore, a simple non-linear model may explain phenomena, which complicated linear models may fail to reproduce. In fact, this generally applies to systems that can show several possible states (i.e. systems which do not have just one stable equilibrium). Out-of-equilibrium models are also required for the description of systemic crises such as the current financial crisis (Helbing, D., 2009, *System risks*).

## 4 Challenges of Socio-Economic Modeling

Many people before and after Popper have been thinking about the logic of scientific discovery (Popper, K. R., 1959). A widespread opinion is that a good model should be applicable to measurements of many systems of a certain kind, in particular to measurements in different areas of the world. The more observations a model can explain and the less parameters it has, the more powerful it is usually considered to be. Models with a few parameters can often be easier to calibrate, and cause-and-effect relationships may be better identified, but one can usually not expect that these models would provide an exact description of reality. Nevertheless, a good model should make predictions regarding some possible, but previously unobserved system behaviors. In this connection, prediction does not necessarily mean the forecast of a certain event at a specific future point in time. It means a specific system behavior that is expected to occur (or to be possible) under certain conditions (e.g. for certain parameter combinations or certain initial conditions). When such conditions apply and the system shows the expected behavior, this would be considered to verify the model, while the model would be falsified or seriously questioned, if the predicted system behavior would not occur. By experimentally challenging models based on their predictions (implications), it has been possible in the natural sciences to rate alterna-

tive models based on their quality in reproducing and predicting measured data. Unfortunately, it turns out that this approach is less suited to identify “the right model” of a social or economic system under consideration. As we will discuss in the following, this is not only due to the smaller amount of data available on most aspects of social and economic systems and due to experimental limitations for financial, technical and ethical reasons.

## 4.1 Promises and Difficulties of the Experimental Approach

So far, it is very expensive to carry out social and economic experiments, for example in the laboratory. While the study of human behavior under controlled conditions has become a common research method not only in psychology, but also in experimental economics and in sociology, the number of individuals that can be studied in such experiments is limited. This implies a large degree of statistical variation, which makes it difficult to determine behavioral laws or to distinguish between different models. The statistical noise creates something like a foggy situation, which makes it difficult to see what is going on. In physics, this problem can be usually solved by better measurement methods (apart from uncertainty that results from the laws of quantum mechanics). In social systems, however, there is an irreducible degree of randomness. The behavior varies not only between individuals due to their heterogeneity (different “personality”). It also varies from one instance to another, i.e. the decision-making of an individual is usually not deterministic. This could be due to various reasons: unknown external influences (details attracting the attention of the individual) or internal factors (exploration behavior, decisions taken by mistake, memory effects, etc.). The large level of behavioral variability within and between individuals is probably not only due to the different histories individuals have, but also due to the fact that exploration behavior and the heterogeneity of behaviors are beneficial for the learning of individuals and for the adaptability of human groups to various environmental conditions. Applying a theory of social evolution would, therefore, suggest that randomness is significant in social and economic systems, because it increases system performance. Besides, heterogeneity can also have individual benefits, as differentiation facilitates specialization. The benefit of a variation between individuals is also well-known from ecological systems (Tilman, D., et al., 1996).

Besides impeding the discovery of behavioral laws, the limited number of participants in laboratory experiments also restricts the number of repetitions and the number of experimental settings or parameter combinations that can be studied. Scanning parameter spaces is impossible so far, while it would be useful



to detect different system behaviors and to determine under which conditions they occur. It can be quite tricky to select suitable system parameters (e.g. the payoff matrix in a game-theoretical experiment). Computer simulations suggest that one would find interesting results mainly, if the parameters selected in different experimental setups imply different system behaviors, i.e. if they belong to different “phases” in the parameter space (see Fig. 1). In order to determine such parameter combinations, it is advised to perform computer simulations before, to determine the phase diagram for the system under consideration (Helbing, D., and Yu, W., 2010). The problem, however, is that the underlying model is unlikely to be perfect, i.e. even a good social or economic model is expected to make predictions which are only approximately valid. As a consequence, the effect one likes to show may appear for (somewhat) different parameter values, or it may not occur at all (considering the level of randomness) (Traulsen, A., et al., 2010).

## 4.2 Several Models Are Right

The above mentioned properties of socio-economic systems imply that it is difficult to select the “right” model among several alternative ones. For an illustration, let us take car-following models, as they are used for the simulation of urban or freeway traffic. Thanks to radar sensors, it has become possible to measure the acceleration of vehicles as a function of the typical variables of car-following models, which are the distance to the car ahead, the own speed, and the speed difference. When fitting the parameters of different car-following models to data of such measurements, it turns out that the remaining error between computer simulations and measurements is about the same for most of the models. The calibration error varies between 12 and 17 percent, and according to the authors, “no model can be denoted to be the best” (Brockfeld, E., et al., 2004). When the error of different models (i.e. the deviation between model and data) is determined for a new data set (using the model parameters determined with the previous data set), the resulting validation error usually varies between 17 and 22 percent (larger validation errors mainly result, when the calibration data set is overfitted) (Brockfeld, E., et al., 2004). Again, the performance of the different models is so similar that it would not be well justified to select one of them as the “correct” model and exclude all the others. A closer analysis shows that the parameters of the car-following dynamics varies among different drivers, but the behavior of specific drivers also varies over time (Kesting, A., and Treiber, M., 2008). We have to assume that the same applies to basically all kinds of behavior, not only for driving a car. Moreover, it is likely that many behaviors (such as decision-making behaviors) vary even more than car-following behavior does. As a consequence,

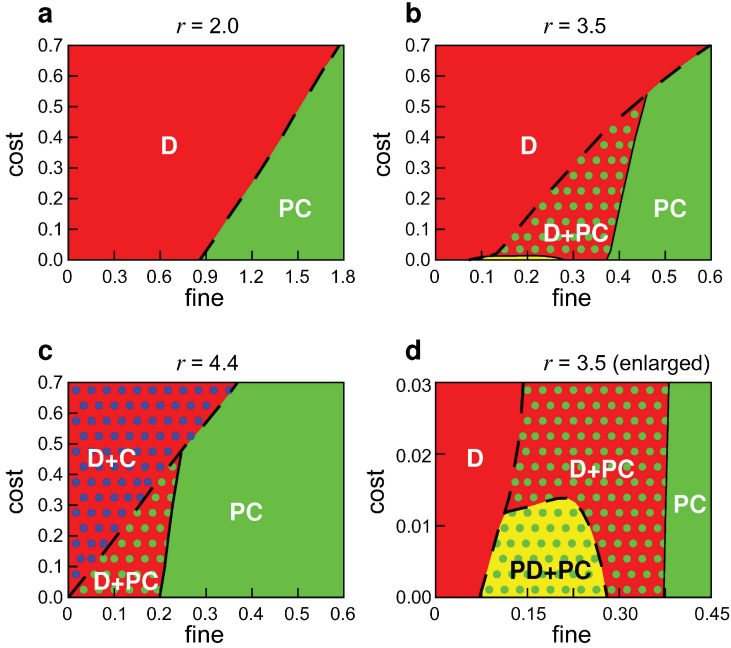
it would be even more difficult to distinguish between different models by means of empirical or experimental data, which would mean that we may have to accept several models to be (possibly) “right”, even when they are not consistent with each other. In other words, the question “Which is the best model?” or “How to choose the model?” may not be decidable in a reasonable way, as is also suggested by the next section. This situation reminds a bit of Gödel’s Undecidability Theorem (Gödel, K., 1962), which relates to the (in)completeness of certain axiom systems.

It may be tempting to determine the best model as the one, which is most successful, for example in terms of the number of citations, it gets. However, success is not necessarily an indicator of a good model. Let us take models used for stock trading as an example. Clearly, even if the stock prices vary in a perfectly random manner and if the average success of each model is the same over an infinite time period; when different traders apply different trading models, they will be differently successful at any chosen point in time. Therefore, one would consider some models more successful than others, while this would be only a matter of luck. As a matter of chance, at other points in time, different models would be the most successful ones.

Of course, if behaviors are not just random so that behavioral laws that go beyond statistical distributions exist, some models should be better than others, and it should eventually be possible to separate “good” from “bad” models through the “wisdom of crowds” effect. However, the “wisdom of crowds” assumes independent judgements, while scientists have repeated interactions. It has been shown experimentally that this tends to create consensus, but that this consensus will often deviate from the truth (Lorenz, J., et al., 2010). The problem results from social influence, which creates a herding effect that can undermine the “wisdom of crowds”. Of course, this mainly applies, when the facts are not sufficiently obvious, which is the case in the social sciences due to the high variability of observations, while the problem is less pressing in the natural sciences thanks to the higher measurement precision. Nevertheless, the physicist Max Planck is known for the quote: “Science progresses funeral by funeral”<sup>4</sup>. Kuhn’s study of scientific revolutions (Kuhn, T. S., 1962) suggests as well that scientific progress is not continuous, but there are sudden paradigm shifts. This reveals the problem of herding effects. Even a collective agreement is no guarantee for the correctness of a model, as the replacement of classical mechanics by relativistic quantum theory shows. In other words, success is not necessarily an indica-

---

<sup>4</sup> Max Planck: “An important scientific innovation rarely makes its way by gradually winning over and converting its opponents, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.”



**Figure 1:** Phase diagram showing the finally remaining strategies in the spatial public goods game with cooperators (C), defectors (D), cooperators who punish defectors (PC) and hypocritical punishers (PD), who punish other defectors while defecting themselves (after Helbing, D., et al., 2010). Initially, each of the four strategies occupies 25% of the sites of the square lattice, and their distribution is uniform in space. However, due to their evolutionary competition, two or three strategies die out after some time. The finally resulting state depends on the punishment cost, the punishment fine, and the synergy  $r$  of cooperation (the factor by which cooperation increases the sum of investments). The displayed phase diagrams are for (a)  $r = 2.0$ , (b)  $r = 3.5$ , and (c)  $r = 4.4$ . (d) Enlargement of the small-cost area for  $r = 3.5$ . Solid separating lines indicate that the resulting fractions of all strategies change continuously with a modification of the punishment cost and punishment fine, while broken lines correspond to discontinuous changes. All diagrams show that cooperators and defectors cannot stop the spreading of costly punishment, if only the fine-to-cost ratio is large enough (see green PC area). Note that, in the absence of defectors, the spreading of punishing cooperators is extremely slow and follows a voter model kind of dynamics. A small level of strategy mutations (which continuously creates a small number of strategies of all kinds, in particular defectors) can largely accelerate the spreading of them. Furthermore, there are parameter regions where punishing cooperators can crowd out "second-order free-riders" (non-punishing cooperators) in the presence of defectors (D+PC). Finally, for low punishment costs, but moderate punishment fines, it may happen that "moralists", who cooperate and punish non-cooperative behavior, can only survive through an "unholy alliance" with "immoral", hypocritical punishers (PD+PC). For related videos, see <http://www.soms.ethz.ch/research/secondorder-free riders> or <http://www.matjazperc.com/games/moral.html>.

tor for good models. It may just indicate which model is most fashionable at a given time. The problem becomes worse by the academic selection process that decides, which scientists make a career and which ones do not. It creates a considerable inertia in the adjustment to new knowledge, i.e. scientific trends are likely to persist longer than what is justified by facts.

### 4.3 No Known Model is Right

A typical approach in the natural sciences is to verify or falsify previously untested predictions (implications) of alternative models by sometimes quite sophisticated experiments. Only in the minority of cases, two alternative theories turn out to be the same, like the wave and the particle picture of quantum mechanics. In most cases, however, two theories A and B are non-identical and inconsistent, which means that they should make different predictions in particular kinds of situations. Experiments are performed to find out whether theory A or theory B is right, or whether both of them deviate from the measurements. If the experimental data confirm theory A and are not compatible with theory B (i.e. deviate significantly from it), one would discard theory B forever. In this way, experiments are thought to narrow down the number of alternative theories, until a single, “true” theory remains.

When social or economic systems are modeled, the following situation is not unlikely to happen: Scientists identify mutually incompatible predictions of theories A and B, and it turns out that an experiment supports theory A, but not theory B. One day, another scientist identifies a different set of incompatible predictions, and another experiment supports theory B, but not theory A. Due to the inherent simplifications of socio-economic models, for any model it should be easy to find empirical evidence that contradicts it. What should one do in such cases? Giving up on modeling would probably not be the best idea. Generalizing a model is always possible, but it will usually end up with detailed models, which imply a number of problems that have been outlined in Sec. 3.2. One could also stay with many particular models and determine their respective ranges of validity. This, however, will never result in a holistic or systemic model. A possible way out would be the approach of pluralistic modeling outlined in Sec. 5.1.

Modeling in modern physics seems to face similar problems. While one would expect that each experiment narrows down the number of remaining, non-falsified models, one actually observes that, after each experiment, scientists come up with a number of new models. As people say: “Each answered question raises ten new ones.” In fact, there is an abundance of elementary particle models, and the same applies to cosmological models. Many models require assuming the

existence of factors that have never been measured and perhaps will never be measured, such as Higgs bosons, dark matter, or dark energy. We will probably have to live with the fact that models are just models that never grasp all details of reality. Moreover, as has been pointed out, understanding elementary particles and fundamental forces in physics would not explain at all what is happening in the world around us (Vicsek, T., 2002; Pietronero, L., 2008). Many emergent phenomena that we observe in the biological, economic and social world will never be derived from elementary particle physics, because emergent properties of a system cannot be understood from the properties of its system components alone. They usually come about by the interaction of a large number of system components. Let us be honest: We do not even understand the particular properties of water, as simple as  $H_2O$  molecules may be.

Generally, there is a serious lack in understanding the connection between function, dynamics, and form. Emergence often seems to have an element of surprise. The medical effect of a new chemical drug cannot be understood by computer simulation alone. So far, we also do not understand emotions and consciousness, and we cannot calculate the biological fitness of a species in the computer. The most exciting open puzzles in science concern such emergent phenomena. It would be interesting to study, whether social and economic phenomena such as trust, solidarity, and economic value can be understood as emergent phenomena as well (Helbing, D., 2010, *Grand socio-economic challenges*).

#### **4.4 The Model Captures Some Features, But May Be Inadequate**

Scientists are often prompted to transfer their methods to another area of application, based on analogies that they see between the behavior of different systems. Systems science is based on such analogies, and physicists generalize their methods as well. The question is how useful a “physicalist approach” can be, which transfers properties of many-particle systems to social or economic systems, although individuals are certainly more intelligent than particles and have many more behavioral degrees of freedom.

Of course, physicists would never claim that particle models could provide an exact description of social or economic systems. Why, then, do they think the models could make a contribution to the understanding of these systems? This is, because they have experience with what can happen in systems characterized by the non-linear interaction of many system components in space and time, and when randomness plays a role. They know how self-organized collective phenomena on the “macroscopic” (aggregate) level can result from interactions

on the “microscopic” (individual) level. And they have learned, how this can lead to phase transitions (also called “regime shifts” or “catastrophes”), when a system parameter (“control parameter”) crosses a critical point (“tipping point”). Furthermore, they have discovered that, at a critical point, the system typically shows a scale-free behavior (i.e. power laws or other fat-tail distributions rather than Gaussian distributions).

It is important to note that the characteristic features of the system at the critical point tend to be “universal”, i.e. they do not depend on the details of the interactions. This is, why physicists think they can abstract from the details. Of course, details are expected to be relevant when the system is not close to a critical point. It should also be added that there are a couple of different kinds of universal behavior, so-called universality classes. Nevertheless, many-particle models may allow one to get a better understanding of regime shifts, which are not so well understood by most established models in economics or the social sciences. However, if the tipping point is far away, the usefulness of many-particle models is limited, and detailed descriptions, as they are favored by economists and social scientists, appear to be more adequate.

Sometimes, it is not so clear how far analogies can carry, or if they are useful at all. Let us take neural network models. In a certain sense, they can be used to model learning, generalization, and abstraction. However, the hope that they would explain the functioning of the brain has been largely disappointed. Today, we know that the brain works quite differently, but neural network theory has given birth to many interesting engineering applications that are even commercially applied. Let us consider models of cooperation based on coupled oscillators as a second example. Without any doubt, the synchronization of cyclical behavior is among the most interesting collective phenomena we know of, and models allow one to study if and how groups of oscillators will coordinate each other or fall apart into subgroups (which are not synchronized among each other, while the oscillators in each of them are) (Mikhailov, A. S., and Calenbuhr, V., 2002). Despite this analogy to group formation and group dynamics, it is not clear, what we can learn from such models for social systems. A similar point is sometimes raised for spin models, which have been proposed to describe opinion formation processes or the emergence of cooperation in social dilemma situations. In this connection, it has been pointed out that social interactions cannot always be broken down into binary interactions. Some interactions involve three or more individuals at the same time, which may change the character of the interaction. Nevertheless, similar phenomena have been studied by overlaying binary interactions, and it is not fully clear how important the difference is.

Let us finally ask whether unrealistic assumptions are generally a sign of bad models. The discussion in Sec. 3.3 suggests that this is not necessarily so. It seems

more a matter of the purpose of a model, which determines the level of simplification, and a matter of the availability of better models, i.e. a matter of competition. Note, however, that a more realistic model is not necessarily more useful. For example, many car-following models are more realistic than fluid-dynamic traffic models, but they are not suited to simulate large-scale traffic networks in real-time. For social systems, there are a number of different modeling approaches as well, including the following:

- Physical(istic) modeling approach: Socio-and econo-physicists often abstract social interactions so much that their models come down to multi-particle models (or even spin models with two behavioral options). Such models focus on the effect of non-linear interactions and are a special case of bounded rationality models, sometimes called zero-intelligence models (Bentley, R. A., and Ormerod, P., *forthcoming*). Nevertheless, they may display features of collective or swarm intelligence (Moussaid, M., et al., 2009). Furthermore, they may be suited to describe regime shifts or situations of routine choice (Gintis, H., 2009), i.e. situations where individuals react to their environment in a more or less subconscious and automatic way. Paul Ormerod, an economist by background, argues as follows (Ormerod, P., 2008): “In many social and economic contexts, self-awareness of agents is of little consequence... No matter how advanced the cognitive abilities of agents in abstract intellectual terms, it is as if they operate with relatively low cognitive ability within the system... The more useful null model in social science agent modelling is one close to zero intelligence. It is only when this fails that more advanced cognition of agents should be considered.”
- Economic modeling approach: Economists seem to have quite the opposite approach. Their concept of “homo economicus” (the “perfect egoist”) assumes that individuals take strategic decisions, choosing the optimal of their behavioral options. This requires individuals with infinite memory and processing capacities. Insofar, one could speak of an infinite-intelligence approach. It is also known as rational choice approach and has the advantage that the expected behaviors of individuals can be axiomatically derived. In this way, it was possible to build the voluminous and impressive theory of mainstream economics. Again, the reliability of this theory depends, of course, on the realism of its underlying assumptions.
- Sociological modeling approach: Certain schools of sociologists use rational choice models as well. In contrast to economists, however, they do not generally assume that individuals would radically optimize their own profit. Their models rather consider that social exchange is more differentiated and multifaceted. For example, when choosing their behavior, individuals may not only consider their own preferences, but the preferences of their interaction

- partner(s) as well. In recent years, “fairness theory” has received a particular attention (Fehr, E., and Schmidt, K. M., 1999) and often been contrasted with rational choice theory. These social aspects of decision-making are now eventually entering economic thinking as well (Frey, B., 1999).
- Psychological modeling approach: Psychologists are perhaps least axiomatic and usually oriented at empirical observations. They have identified behavioral paradoxies, which are inconsistent with rational choice theory, at least its classical variant. For example, it turns out that most people behave in a risk averse way. To account for their observations, new concepts have been developed, including prospect theory (Kahneman, D., and Tversky, A., 1979), satisficing theory (Simon, H. A., 1955), and the concept of behavioral heuristics (Gigerenzer, G., et al., 2000). In particular, it turns out that individual decisions depend on the respective framing. In his economic Nobel lecture, Daniel Kahneman put it this way: “Rational models are psychologically unrealistic... the central characteristic of agents is not that they reason poorly, but that they often act intuitively. And the behavior of these agents is not guided by what they are able to compute, but by what they happen to see at a given moment.” Therefore, modern research directions relate to the cognitive and neurosciences. These results are now finding their way into economics via the fields of experimental, behavioral, and neuro-economics.

In summary, there is currently no unified approach that scientists generally agree on. Some of the approaches are more stylized or axiomatic. Others are in better quantitative agreement with empirical or experimental evidence, but mathematically less elaborated. Therefore, they are theoretically less suited to derive implications for the behavior in situations, which have not been explored so far. Consequently, all models have their strengths and weaknesses, no matter how realistic they may be. Moreover, none of the mathematical models available so far seems to be sophisticated enough to reflect the full complexity of social interactions between many people.

## 5 Discussion and Outlook

### 5.1 Pluralistic or Possibilistic Modeling and Multiple World Views: The Way Out?

Summarizing the previous discussion, it is quite unlikely that we will ever have a single, consistent, complete, and correct model of socio-economic systems.



Maybe we will not even find such a grand unified theory in physics. Recently, doubts along these lines have even been raised by some particle physicists (Woit, P., 2006; Smolin, L., 2007). It may be the time to say good-bye to a modeling philosophy that believes in the feasibility of a unique, general, integrated and consistent model. At least there is no theoretical or empirical evidence for the possibility of it.

This calls for a paradigm shift in the modeling approach. It is important to be honest that each model is limited, but most models are useful for something. In other words, we should be tolerant with regard to each other's models and see where they can complement each other. This does not mean that there would be separate models for non-overlapping parts of the system, one for each subsystem. As has been pointed out, it is hard to decide whether a particular model is valid, no matter how small the subsystem is chosen. It makes more sense to assume that each model has a certain validity or usefulness between 0 and 1, and that the validity furthermore depends on the part or aspect of the system addressed. This validity may be quantified, for example, by the goodness of fit of a given system or the accuracy of description of another system of the same kind. As there are often several models for each part or aspect of a system, one could weight the models with their respective validity, as determined statistically. Analogously to the "wisdom of crowds" (Galton, F., 1907), which is based on the law of large numbers, this should lead to a better quantitative fit or prediction than most (or even each) model in separation, despite the likely inconsistency among the models. Such an approach could be called a pluralistic modeling approach (Rotmans, J., and Asselt, M. B. A. van, 2001), as it tolerates and integrates multiple worldviews. It may also be called a possibilistic approach (Dubois, D., and Prade, H., 2004), because it takes into account that each model has only certain likelihood to be valid, i.e. each model describes a possible truth. However, this should not be misunderstood as an appeal for a subjectivistic approach. The pluralistic modeling approach still assumes that there is some underlying reality that some, many, or all of us share (depending on the aspect we talk about).

As shocking as it may be for many scientists and decision-makers to abandon their belief in the existence of a unique, true model, the pluralistic modeling approach is already being used. Hurricane prediction and climate modeling are such examples (Lucarini, V., 2002). Even modern airplanes are controlled by multiple computer programs that are run in parallel. If they do not agree with each other, a majority decision is taken and implemented. Although this seems pretty scary, this approach has worked surprisingly well so far. Moreover, when crash tests of newly developed cars are simulated in the computer, the simulations are again performed with several models, each of which is based on different approximation methods.

It is plausible to assume that pluralistic modeling will be much more widely used in future, whenever a complex system shall be modeled.

## 5.2 Where Social Scientists and Natural Scientists or Engineers Can Learn From Each Other

It has been argued that each modeling approach has its strength and weaknesses, and that they should be considered complementary rather than competitive. This also implies that scientists of different disciplines may profit and learn from each other. Areas of fruitful multi-disciplinary collaboration could be:

- the modeling of socio-economic systems themselves,
- understanding the impacts that engineered systems have on the socio-economic world,
- the modeling of the social mechanisms that drive the evolution and spreading of innovations, norms, technologies, products etc.,
- scientific challenges relating to the managing of complexity and to systems design,
- the application of social coordination and cooperation mechanisms to the creation of self-organizing technical systems (such as decentralized traffic controls or peer-to-peer systems),
- the development of techno-social systems (Vespignani, A., 2009), in which the use of technology is combined with social competence and human knowledge (such as Wikipedia, prediction markets, recommender systems, or the semantic web).

Given the large potentials of such collaborations, it is time to overcome disciplinary boundaries. They seem to make less and less sense. It rather appears that multi-disciplinary, large-scale efforts are needed to describe and understand socio-economic systems well enough to address practical challenges of humanity (such as the financial and economic crisis) more successfully (Helbing, D., 2010, *The FuturICT knowledge*).

## References

- Albeverio, S., Jentsch, V., & Kantz, H. (eds.) (2005). *Extreme Events in Nature and Society*. Berlin: Springer.
- Bentley, R. A. & Ormerod, P. (forthcoming). *Agents, intelligence, and social atoms*. Preprint available at <http://www.paulormerod.com/wp-content/uploads/2012/06/agents.pdf>.

- Bertalanffy, L. von (1968). *General System Theory: Foundations, Development, Applications*. New York: George Braziller.
- Bollinger, Lee C. (2005). Announcing the Columbia Committee on Global Thought. See <http://www.columbia.edu/cu/president/docs/communications/2005-2006/051214-committee-global-thought.html>.
- Box, G. E. P. & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. New York: John Wiley and Sons. 74, 424.
- Brockfeld, E., Kühne, R. D., & Wagner, P. (2004). Calibration and validation of microscopic traffic flow models. *Transportation Research Board* 1876. 62–70.
- Comte, A. (1830–1842). *Course on Positive Philosophy*.
- Comte, A. (1856). *Social Physics: From the Positive Philosophy*. New York: Calvin Blanchard.
- Davidson, P. A. (2004). *Turbulence*. Cambridge: Cambridge University Press.
- Dyson, F. (2004). A meeting with Enrico Fermi. *Nature* 427. 297.
- Dubois, D. & Prade, H. (2004). Possibilistic logic: a retrospective and prospective view. *Fuzzy Sets and Systems* 144(1). 3–23.
- Einstein, A. (1934). On the Method of Theoretical Physics. *The Herbert Spencer Lecture. Philosophy of Science* 1(2). Delivered at Oxford (10 June 1933). 165.
- Epstein, J. M. (2006). *Generative Social Science. Studies in Agent-Based Computational Modeling*. Princeton, NJ: Princeton University Press. 51.
- Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation* 11(4). 12. See <http://jasss.soc.surrey.ac.uk/11/4/12.html>.
- Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114(3). 817–868.
- Floreano, D. & Mattiussi, C. (2008). *Bio-Inspired Artificial Intelligence: Theories, Methods, and Technologies*. Cambridge, MA: MIT Press.
- Frey, B. (1999). *Economics as a Science of Human Behaviour: Towards a New Social Science Paradigm*. Dordrecht: Kluwer Academics.
- Galton, F. (1907). Vox populi. *Nature* 75. 450–451.
- Gigerenzer, G., Todd, P. M., & ABC Research Group (2000). *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.
- Gilbert, N. (ed.) (2010). *Computational Social Science*. Los Angeles: Sage.
- Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton: Princeton University Press.
- Gödel, K. (1962). *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. New York: Basic.
- Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.
- Helbing, D. (2001). Traffic and related self-driven many-particle systems. *Reviews of Modern Physics* 73. 1067–1141.
- Helbing, D. (2009). Derivation of non-local macroscopic traffic equations and consistent traffic pressures from microscopic car-following models. *European Physical Journal B* 69(4). 539–548. See also <http://www.soms.ethz.ch/research/traffictheory>.
- Helbing, D., Deutsch, A., Diez, S., Peters, K., Kalaidzidis, Y., Padberg, K., Lämmer, S., Johansson, A., Breier, G., Schulze, F., & Zerial, M. (2009). Biologistics and the struggle for efficiency: Concepts and perspectives. *Advances in Complex Systems* 12(6). 533–548.
- Helbing, D. (2009). System risks in society and economics. Santa Fe Institute Working Paper #09–12–044. See <http://www.santafe.edu/media/workingpapers/09-12-044.pdf>.

- Helbing, D., Szolnoki, A., Perc, M., & Szabó, G. (2010). Evolutionary establishment of moral and double moral standards through spatial interactions. *PLoS Computational Biology* 6(4). e1000758.
- Helbing, D. & Yu, W. (2010). The future of social experimenting. *Proceedings of the National Academy of Sciences USA (PNAS)* 107 (12). 5265–5266. See also <http://www.soms.ethz.ch/research/socialexperimenting>.
- Helbing, D. (2010). *The FuturICT knowledge accelerator: Unleashing the power of information for a sustainable future*. Project Proposal. See <http://arxiv.org/abs/1004.496>.
- Helbing, D. (2010). *Grand socio-economic challenges*. Working Paper, ETH Zurich.
- Horsthemke, W. & Lefever, R. (1983), *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry, and Biology*. Berlin: Springer.
- Jackson, M. O. (2008). *Social and Economic Networks*. Princeton, NJ: Princeton University Press.
- Kagel, J. H. & Roth, A. E. (1995). *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47(2). 263–291.
- Kesting, A. & Treiber, M. (2008). Calibrating car-following models by using trajectory data: Methodological study. *Transportation Research Record* 2088. 148–156.
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2010) *How social influence undermines the wisdom of crowds*. Submitted.
- Lucarini, V. (2002). Towards a definition of climate science. *International Journal of Environment and Pollution* 18(5). 413–422.
- Maimon, O. & Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook*. New York: Springer.
- Mayergoyz, I. D. (2003). *Mathematical Models of Hysteresis and their Applications*. New York: Academic Press.
- Mikhailov, A. S. & Calenbuhr, V. (2002). *From Cells to Societies. Models of Complex Coherent Action*. Berlin: Springer.
- Miller, J. H. & Page, S. E. (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton, NJ: Princeton University Press.
- Moussaid, M., Garnier, S., Theraulaz, G., & Helbing, D. (2009). Collective information processing and pattern formation in swarms, flocks, and crowds. *Topics in Cognitive Science* 1(3). 469–497.
- Nolfi, S. & Floreano, D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. Cambridge, MA: MIT Press.
- Ormerod, P. (2008). What can agents know? The feasibility of advanced cognition in social and economic systems. Communication, Interaction and Social Intelligence. In: *Proceedings of the AISB 2008 convention: Communication, interaction and social intelligence*. Aberdeen. See <http://www.paulormerod.com/wp-content/uploads/2012/06/Whatcanagentsknow.pdf>.
- Pietronero, L. (2008). Complexity ideas from condensed matter and statistical physics. *europhysics news* 39(6). 26–29.
- Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson. Original German version: Popper, K. R. (1935). *Logik der Forschung*. Vienna: Springer.
- Rotmans, J. & Asselt, M. B. A. van (2001). Uncertainty management in integrated assessment modeling: Towards a pluralistic approach. *Environmental Monitoring and Assessment* 69(2). 101–130.

- Schuster, H. G. & Just, W. (2005). *Deterministic Chaos*. Weinheim: Wiley-VCH.
- Schweitzer, F. (ed.) (1997). *Self-Organization of Complex Structures: From Individual to Collective Dynamics*. London: Gordon and Breach.
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1). 99–118.
- Smolin, L. (2007). *The Trouble With Physics: The Rise of String Theory, The Fall of a Science, and What Comes Next*. Boston: Mariner.
- Sornette, D. (2006). *Critical Phenomena in Natural Sciences. Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*. Berlin: Springer.
- Spencer, H. (1898). *The Principles of Sociology*. New York: Appleton. (The three volumes were originally published in serial form between 1874 and 1896).
- Stanley, H. E. (1987). *Introduction to Phase Transitions and Critical Phenomena*. Oxford: Oxford University Press.
- Tilman, D., Wedin, D., & Knops, J. (1996). Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* 379. 718–720.
- Traulsen, A., Semmann, D., Sommerfeld, R. D., Krambeck, H.-J., & Milinski, M. (2010). Human strategy updating in evolutionary games. *Proceedings of the National Academy of Sciences USA (PNAS)* 107(7). 2962–2966.
- Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science* 325. 425–428.
- Vicsek, T. (2002). The bigger picture. *Nature* 418. 131.
- Weidlich, W. (2006). *Sociodynamics: A Systemic Approach to Mathematical Modelling in the Social Sciences*. London: Dover.
- Woit, P. (2006). *Not Even Wrong: The Failure of String Theory and the Search for Unity in Physical Law*. New York: Basic.
- Zeeman, E. C. (ed.) (1977). *Catastrophe Theory*. London: Addison-Wesley.

### **Dieser Beitrag wurde bereits online veröffentlicht:**

Helbing, D. (2010). Pluralistic modeling of complex systems.  
*Science and Culture* 76 (9–10), 315–329.  
[http://scienceandculture-isna.org/sept\\_oct\\_10/03%20Dirk%20Helbing.pdf](http://scienceandculture-isna.org/sept_oct_10/03%20Dirk%20Helbing.pdf)

### **Prof. Dr. Dirk Helbing**

ETH Zurich  
 CLU E 1  
 Clausiusstr. 50  
 8092 Zürich  
 Switzerland  
[dhelbing@ethz.ch](mailto:dhelbing@ethz.ch)



Stephan Hartmann

# **The Methodological Challenges of Complex Systems**

Commentary on Dirk Helbing

## **1 Introduction**

In this rich and insightful paper, Dirk Helbing addresses a large number of methodological problems regarding the modeling of complex systems. Some of these problems concern models of complex systems in general; others concern particular models of socio-economic systems. The latter are the most complex systems we can think of, and it is by no means clear (and accordingly controversial) whether the modeling approach can succeed here at all. However, there has been a lot of progress in this field in the last years. This progress is made possible by the increase in computer power and the highly interdisciplinary nature of this research, which led to a variety of different modeling approaches. Helbing has tremendously illuminating things to say about all this, and I recommend this paper highly to all philosophers of science interested in modeling.

In this short commentary, I will discuss two central methodological points that Helbing makes: his plea for pluralistic modeling (Sec. 2) and the possibilistic approach that he advocates (Sec. 3).

## **2 Pluralistic Modeling**

Helbing explains that we find a whole spectrum of types of models in the new field of socio-economic modeling. While some of these models are more detailed, others are simple (so-called ‘toy models’). Whereas some models use a physical approach, others use an economic, sociological, or psychological approach. It is important to note that all types of models have various merits and shortcomings. Even if we consider different models of the same type, we find that some models work well for a certain class of applications, while other models work well for other applications. No single model gets everything right, which makes it hard or even impossible to identify the one true model (if there is one). Besides, different models have different functions. For example, while some models allow us to make accurate predictions, others give us insight and understanding. Interest-

ingly, no single model fulfills all desired functions. For these reasons, Helbing advocates a pluralistic approach to socio-economic modeling according to which the scientific community should study a variety of models for a certain phenomenon even if the different models are not consistent with each other and presuppose different “world views”. Here are a few remarks on these claims:

I think that Helbing is right to stress that different models have different functions and that the scientific community values and desires all these functions (Frigg and Hartmann 2006). However, it may happen that different scientists give different weights to a certain function. For example, one scientist might find it most important that a model makes accurate predictions. Such a scientist will presumably favor a more detailed model to a simple model, as more detailed models typically lead to better predictions. Another scientist, who is interested in gaining understanding, will prefer a simple model which is easier to grasp, and from which she hopes to identify the essential features that bring about the phenomenon under consideration. It seems that both scientists are rational in their choice. Consequently, there does not seem to be a unique model choice, which is independent of the (arguably equally rational) epistemic preferences of the scientists. However, if two scientists agree on their epistemic preferences (i.e. on the weights they assign to different functions), they might well (and perhaps even *should*) agree on their model choice.

Our previous discussion has presupposed that the different functions of models are independent and mutually irreducible. That is, we assume, for example, that a predictively successful model does not automatically also provide good explanations. Moreover, the different functions are not entailed by a common goal such as truth. That is, we follow Cartwright (1983), who famously argued that “the truth does not explain much”. Cartwright rightly observed that simple explanatory models are often far from empirically adequate, while detailed models do not provide much insight (Hartmann 1998). And so we have to make a choice. This typically means that some members of the scientific community explore simple models, while others explore more complicated models. Again, others explore some models in-between.

Despite Cartwright’s skepticism, it is controversial amongst philosophers of science whether, and to what extent, the various functions of models can be reduced to the goal of truth. For example, some authors have presented ingenious arguments to show that unification (Myrvold 2003) and simplicity (Swinburne 1997) are truth-conducive. And while a lot of progress has been made here, not everybody is convinced. Consequently, I assume that a number of different epistemic and pragmatic values are associated with scientific theories and models, which we take to be irreducible, to the effect that different scientists (or different parts of the scientific community) endorse different values. There is another



reason why the goal of truth is not privileged. After all, why would the scientific community care about a true but otherwise useless (e.g. too complicated) model?

To proceed, let us assume that two scientists agree on their epistemic preferences and ask whether a proliferation of models is methodologically advisable. Two considerations come to mind here: First, one may say that a proliferation of models is advantageous as the availability of alternatives helps finding the best model that satisfies the agreed-upon epistemic preferences. The belief here is that there is a best model, and that it is only a matter of effort to find it. Exploring several alternative models at the same time will then speed up the process of arriving at the best model. This is the *epistemically optimistic* view associated pluralism. Alternatively, one may believe that there is no best model, and that all we can do is to entertain a number of alternative models, explore them, and apply them as good as we can. This is the *epistemically pessimistic* view of pluralism.

It is not quite clear which of the two options Helbing favors. He presents some polemics against the ‘one-true-model’ view, but I am not sure whether his arguments also apply if we agree on the functions (and their respective weights). Helbing also does not say much about the conditions of adequacy of an acceptable model. Are all models equally acceptable? Or are there certain conditions that a model has to satisfy to be acceptable or (at least) entertainable? Does a model have to score high on at least one function? And: What if a model conflicts with some data? Shall we then wait and see if it accounts well for other data, or shall we reject the model right away in this case? It is difficult to find answers to these questions and to formulate reliable criteria. Much seems to depend on the details of the specific case, and on the judgments of the respective scientists.

Let us turn to the relation between different models for the same phenomenon. Helbing observes that different models of the same phenomenon make different assumptions about the world. Thus, it is interesting to ask whether these assumptions (and hence, the models) are compatible with each other. Helbing seems to be content with different models being inconsistent with each other. I have two comments on this: First, the situation in socio-economic modeling is not at all special in this regard. We also find alternative and seemingly incompatible models in more traditional parts of science, such as nuclear physics. There, we have a similarly rich spectrum of models ranging from, e.g., the liquid drop model to the various shell models. These models make fundamentally different assumptions about their target system and appear to be contradictory. However, there does not seem to be a reason for concern about these apparent inconsistencies. Firstly, all nuclear models are just models and, as such, involve idealizations. They are strictly speaking false and do not tell us the whole truth about the object or system under consideration. The different models rather complement each other, and each model provides (metaphorically speaking) a certain

perspective on the phenomenon in question (Giere 2006). It is also interesting to note that the various models can often be approximately derived from (or at least be made plausible on the basis of) a more fundamental theory (Hartmann 1999, cf. Morrison 2011). Besides, the different models often have different domains of applicability. One model may explain certain aspects of a phenomenon, another model may explain others. If one adds the domain of applicability of a model in a *ceteris paribus* clause, the apparent inconsistency of different models disappears. And so I conclude that we should not worry about apparent inconsistencies. They only become a problem if we take the models too seriously (Hartmann 1996).

### 3 Possibilistic Modeling

Many of my above claims may be too liberal. Is it really false to assume that truth plays a privileged role in scientific theorizing? Should we not instead require that a model is at least approximately true? This seems to be a plausible requirement, as giving up truth altogether and focusing only on the pragmatic functions of models does not seem to do justice to the scientific endeavor. It seems reasonable to think that a model is only possible if it is approximately true, and so Helbing's (not worked out) possibilistic approach seems to require an explication of the notion of 'approximate truth'. To do so, several proposals have been made in the literature, see Festa et al. (2005) and Niiniluoto (1999).

On a related note, it is also hard to assign non-vanishing probabilities to models if we accept that the latter involve idealizations (i.e. false claims). Should we then not assign a probability of zero to the model? And if we do not assign a probability of zero to such a model, what does the probability assignment actually mean? Does it measure the usefulness of the model? But if it does, it is not clear why these (effective) utilities should follow the axioms of probability theory.

These are important questions, which any Bayesian philosopher of science who wants to account for the practice of science (in which models play an important role) has to address. Here, we can only sketch a proposal for how non-vanishing probabilities can be assigned to a given model. The basic idea is that idealization-involving models may nevertheless help us to make good predictions and account for given data *within a certain margin of error*. It seems that, given such an error margin, it does not matter whether we use a highly idealized assumption (such as "the Earth is a point mass"), or a more realistic assumption (i.e. that the Earth has a certain shape and mass distribution) if we want to calculate, e.g., how long a rock needs to fall down from a certain height. Replacing the true assumption by an idealized assumption is justified in this case. Clearly, this idea needs to

be made more precise. But if we do so, it appears possible to assign probabilities to idealized models for a specific application, and given a certain margin of error.

With these probabilities, along with epistemic utilities that weigh the different functions of a model, an individual scientist (or a group of scientists) can then calculate the expected utility of different models and choose the one which maximizes expected utility. This scientist (or group of scientists) will then explore the model further, apply it, and study its domain of applicability. Given the provisional nature of the various models, it is however important that other members of the scientific community focus on other models. And if predictions are expected from the scientific community (as, for example, in the case of climate models), some average of the predictions of the various models should be chosen. But which average? One option is to simply use the straight average, i.e. to give all models the same weight. As Professor Helbing stresses, this strategy uses “the wisdom of the crowds” and often leads to much more reliable predictions. Alternatively, one could weigh each model prediction with its validity (as Professor Helbing suggests). However, I doubt that this works. Firstly, there will not be a consensus on these validities across the scientific community. All we have is subjective assessments of the true validity value. Secondly, to reach a consensus on the validity value, some kind of deliberation has to take place. However, Professor Helbing himself has shown that this procedure often leads us away from the truth (Lorenz et al. 2011). And so I think that taking the straight average of the various model predictions is the best strategy. It is also very easy to implement.

## References

- Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.
- Festa, R., Aliseda, A., & Peijnenburg, J. (eds.) (2005). *Confirmation, Empirical Progress, and Truth Approximation: Essays in Debate with Theo Kuipers*. Amsterdam: Rodopi.
- Frigg, R. & Hartmann, S. (2006). Models in Science. In: *The Stanford Encyclopedia of Philosophy* (Spring 2006 Edition).
- Giere, R. (2006). *Scientific Perspectivism*. Chicago: University of Chicago Press.
- Hartmann, S. (1999). Models and Stories in Hadron Physics. In: Morgan, M. & Morrison, M. (eds.). *Models as Mediators*, Cambridge: Cambridge University Press. 326–346.
- Hartmann, S. (1998). Idealization in Quantum Field Theory. In: Shanks, N. (ed.), *Idealization in Contemporary Physics*, Amsterdam: Rodopi. 99–122.
- Hartmann, S. (1995). Models as a Tool for Theory Construction: Some Strategies of Preliminary Physics. In: Herfel, W. et al. (eds.), *Theories and Models in Scientific Processes*. Amsterdam: Rodopi. 49–67.
- Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How Social Influence Can Undermine the Wisdom of Crowd Effect. *Proceedings of the National Academy of Sciences* 108(20).

- Morrison, M. (2011). One Phenomenon, Many Models: Inconsistency and Complementarity. *Studies in History and Philosophy of Science* 42 (2). 342–351.
- Myrvold, W. (2003). A Bayesian Account of the Virtue of Unification. *Philosophy of Science* 70. 399–423.
- Niiniluoto, I. (1999). *Critical Scientific Realism*. Oxford: Oxford University Press.
- Swinburne, R. (1997). *Simplicity As Evidence of Truth*. Milwaukee: Marquette University Press.

**Prof. Dr. Stephan Hartmann**

LMU Munich  
Munich Center for Mathematical Philosophy  
Geschwister-Scholl-Platz 1  
80539 München  
Germany  
s.hartmann@lmu.de

Uskali Mäki

# Contested Modeling: The Case of Economics

## 1 Introduction

Economics is a culturally and politically powerful and contested discipline, and it has been that way as long as it has existed. For some commentators, economics is the “queen of the social sciences”, while others view it as a “dismal science” (and both of these epithets allow for diverse interpretations; see Mäki 2002). Economics is also a discipline that deals with a dynamically complex subject matter and has a tradition of reducing this complexity by using systematic procedures of simplification. Nowadays, these procedures involve for the most part building and using mathematical models (for an overview of the philosophical issues, see Morgan and Knuuttila 2011). In the dominant circles of the discipline, one is not regarded as a serious economist having a professional expert view on any given economic or social issue without having a model about it. Much of the power of the discipline and its characteristic contestations therefore involve models and modeling: the successes and failures of the dismal queen are those of modeling. The issues involved in economic modeling have been made particularly acute once again by the financial crisis of 2008–2009 and its aftermath: the discipline of economics is among the candidates for the major blame for failure.

I will first outline some thoughts about the characteristic disciplinary conventions that guide and constrain modeling in economics. I will then summarize my account of the very ideas of models and modeling. Finally, within the framework of that account, I will highlight some major issues of contestation and sketch the respective notions of potential success and failure in economic modeling with illustrations. These notions are motivated by my subscription to a (flexible and discipline-sensitive) realist philosophy of science (e.g. Mäki 2005).

## 2 Some characteristics of economics and economic modeling

The notions of model and modeling can be conceived broadly or more narrowly. Broadly conceived, they encompass activities such as theoretical modeling, laboratory experimentation, and computer simulation. These are examples of *surrogate reasoning* that share the strategy of using one thing (the surrogate object) to learn about another thing (the target object). While in other ways different

from one another, the study of theoretical models, laboratory experiments and simulations is similar in being species of surrogate reasoning (for the similarities and differences between theoretical modeling and experimentation, see Morgan 2003; and Mäki 1992, 2005). More narrowly conceived, of these three categories, only theoretical modeling qualifies as proper modeling. The present paper has a narrow focus on theoretical models and modeling in economics.

The prevalence of modeling, whether more broadly or more narrowly conceived, is particularly salient in research fields that seek [a] to access targets that are complex (such as biology, ecology, climatology); [b] to access targets that are distant in time or space, very small, very large, very slow, very fast, or ethically awkward (such as cosmology, archaeology, evolutionary theory, nuclear physics, biomedicine); and [c] to access familiar but complex targets whose overall functioning is often unapparent, possibly for reasons such as those listed above in [a] and [b] (of this, economics is a prominent example).

Each discipline or type of discipline has its own ways – styles, routines, values, and conventions – of modeling, and they are not a simple function of the specific nature of its target domain. It is always somewhat risky to make generalizing claims about a discipline and its characteristic practices and the values guiding those practices, but I believe the following will be recognized as more or less accurate regarding much of economic modeling in the recent decades.

*First*, economic modeling is often theory-driven, shaped and constrained strongly by the dominant theoretical framework. This framework nowadays usually requires that models be built in terms of optimizing agents and equilibrium outcomes.

*Second*, the combination of parsimony and breadth is highly valued in economics. This means that there is an urge to increase the unification of diverse phenomena in terms of portable model structures or modeling principles, that is, structures or principles that can easily be transferred from one domain (or even discipline) to another (see Mäki 2001, 2009).

*Third*, what is typically highly valued in economic modeling is mathematical rather than numerical precision. It is not surprising therefore that analytical derivation tends to be preferred to computer simulation (see Lehtinen and Kuorikoski 2007).

*Fourth*, among the achievements of theoretical modeling economists often mention that models provide some “insight” into phenomena and the mechanisms which produce them; that they yield conditional predictions that state that if certain conditions were to prevail, then this or that would happen; and that they suggest how-possibly explanations that give account of ways in which some given phenomena might have come about (in contrast to how they actually did come about).

*Fifth*, the so-called Duhem-Quine problem of underdetermination of theory or model choice by the empirical data is particularly pressing in economics. In practice this often means that theoretical disputes are hard to settle by empirical means, and that theories and models that were thought to have been refuted by empirical evidence often make a comeback and enjoy long academic lives.

*Sixth*, the dominant streams of 20<sup>th</sup> century economics have for the most part been characterized by one-way disciplinary autonomy, that is, the relative reluctance to import to economics substantive ideas from other disciplines such as sociology or psychology.

Disciplinary conventions are not carved in stone, so they may occasionally be subject to modification or rejection. Some of the above disciplinary conventions of economics are being increasingly questioned and alternatives are being tried out. For example, the proportion of data-driven modeling and computer simulations, even if still relatively small, has been increasing, and there is growing interdisciplinary traffic flowing to economics from experimental psychology and cognitive neurosciences. In the aftermath of the crisis of 2008–2009, many economists have proposed that, in order to understand the mechanisms that tend to bring about this sort of crisis, economists should better do agent-based simulations (e.g. Farmer and Foley 2009) and incorporate “animal spirits” in their models, informed by cognitive sciences broadly conceived (e.g. Akerlof and Shiller 2009). Predicting the future of a discipline is always difficult, but economics may have started becoming more diverse in its disciplinary conventions than has been the case in the recent decades.

Whatever the future of the discipline may hold, its past has had one perennial methodological issue above others. This is the concern of unrealistic models and their assumptions – such as perfect competition, the fully informed self-seeking rational *homo economicus*, instantaneous and cost-free market adjustment in an institutional vacuum, international trade with two countries, two goods and two factors of production, and so on. Within and around economics, nothing compares to the most important methodological issue: what if any justification might be available to unrealisticness in models and their assumptions?

In dealing with this issue, economics have to overcome pressures and worries from two directions. There is the “phenomenological” pressure and the respective worry: *Does the world look like that?* This is the puzzlement among audiences such as beginning economics students and other uninitiated observers such as other social scientists. There is also the “practical” pressure and the related worry: *Does the model work?* This reflects the expectations among the policy-advising economists as well as consumers of economics such as policy makers and the general public as spectators of the performance of economics regarding its policy relevance, akin to other technologically oriented engineering disciplines.

The *phenomenological worry* derives from the fact that ordinary people, including students of economics, are also economic agents with amassed collectively shared experience and commonsense conceptions about the economy. The contrast between theoretical models and phenomenology is often stark. This discrepancy has two very different sources. First, economic models are formulated in terms of assumptions that radically idealize items in the commonsense experience, e.g. when the behaviour of ordinary people is portrayed as that of the fictional *homo economicus*. Second, economic models typically provide (invisible-hand) explanations, which are surprising and counterintuitive from the commonsense point of view, e.g. when free trade is modelled as benefiting all parties and when apparently irrational herd behaviour is modelled as arising from individual rationality.

The *practical worry* is equally pressing. Economics is regularly faced with charges of practical failures. This is an ongoing concern, but in every few decades the credentials of economics are questioned more seriously in public. In these situations, the challenge of academic accountability of the discipline is turned into one of broader public accountability. In fact, we have such a situation right now. On 16<sup>th</sup> July 2009, the *Economist* magazine wrote: “Of all the economic bubbles that have been pricked, few have burst more spectacularly than the reputation of economics itself.”

The two worries – the phenomenological and the practical – imply that a commentator of economic models has to meet special challenges. The clash of theoretical models with commonsense views implies a need to understand the origins of the clash and the associated attitudes and arguments, including attempts to justify theoretical models not only as unproblematic, but also as superior to the commonsense conceptions. The successes and failures of economics in guiding economic policy likewise give rise to the call for explanation and justification of the varying practical performance of the discipline. For these purposes we need accounts of criticism and defence of, as well as success and failure in modeling. They are accounts of various kinds of contestation faced by economic modeling. Before these will be discussed, an account of modeling is needed.

### 3 The very ideas of model and modeling

The key idea of modeling is to examine one thing (the target) by examining another (the model). Using models is motivated by the circumstance that there is no direct and easy epistemic access to the target. A model is, at best, a surrogate object in the following way: By directly examining what happens in the surrogate



object the investigator seeks to indirectly acquire information about the target object. The surrogate object is taken to stand for the target and must be required to be sufficiently similar with it for such information acquisition to be possible. Another way of putting this is saying the surrogate object *represents* the target object.

In many contexts, we are inclined to talk about models in simple dyadic terms such that one thing is a model of another. In order to understand what a model is, however, it is not enough to think of it in terms of a dyadic relation between two objects, the model and its target. Recent philosophical work has stressed the roles of two further components in constituting a model, namely an agent and a purpose: an agent considers or uses one object as a model of another object for some purpose (e.g. Giere 1999). The recent literature has also investigated the notion of representation in connection to models: indeed, models are typically conceived as representations of their targets. Yet there is no elaborate notion of model representation available that would express a consensus view. In my opinion adequate accounts of model and representation should be richer than has been customary. Further elements are needed in addition to agents and purposes associated with models. I have suggested such a richer idea of model representation that takes representation to have two aspects, those of representative and resemblance. It distinguishes between a model and its description, and it adds the ideas of audience and commentary to the overall notion (e.g. Mäki 2009a,b, 2011). Here is a formulation of this account of model representation:

### [ModRep]

Agent *A*

uses (imagined) object *M* as

a **representative** of (actual or possible) target *R*

for **purpose** *P*,

addressing **audience** *E*,

at least potentially prompting genuine **issues of resemblance** between *M* and *R* to arise,

describing *M* and drawing inferences about *M* and *R* in terms of one or more **model descriptions** *D*,

and applies **commentary** *C* to identify and coordinate the other components.

There are several noteworthy features in [ModRep]. Nothing is a model without being used as such by some individual or collective *agent*. Use implies purpose. A model can be used for a variety of different *purposes*, such as predicting some future event or property with a certain degree of accuracy; isolating a fragment of a causal structure; exploring possible causal configurations; serving as a bench-

mark; refining a mathematical technique; designing a well-functioning institution; and so on. Reflecting the social nature of scientific inquiry, models are used in relation to various *audiences* – such as specialists in the same research field, students, policy makers, the curious general public – in order to pursue goals such as communicating information, teaching undergraduate students the core principles of conventional economic reasoning, and persuading some relevant audience to adopt a point of view. The choice of *model description* typically reflects the presumed expectations and competencies of the relevant audience. For example, advanced mathematical languages may be used when addressing expert scientists in the same field, while familiar metaphors and visualizations of various kinds may be relatively more effective when addressing beginning students and lay audiences.

The very idea of models as representations can be briefly summarized. I take representation to involve two aspects: that of representative and that of resemblance. A model represents (rather: is used to represent) a target by being (used as) its *representative*, by standing for it. This is the relatively more voluntary side of modeling: the modeler chooses (or chooses to build) the object that is then used as a representative of some target. This is not yet sufficient for representation: not just any object can reasonably represent the target object. Some further conditions or constraints must be met, and these are not entirely subject to the decision of the modeler. The key condition in [ModRep] is given by the second aspect of representation, that of *resemblance* or similarity with some target. This idea comes with two important qualifications. First, it is not required that the model actually does resemble the target, it is rather required that the model has *a chance of resembling* some target and that this *potentially prompts the issue of whether indeed the model does resemble*. Further inquiry may then settle this issue, but such inquiry is not required for establishing whether the object represents or not. Further inquiry is required for establishing whether the model *truthfully* represents.

Resemblance is a matter of how the model world – the world envisaged in a model – is related to the real world. The second qualification is that what really matters in modeling is not resemblance per se, but rather *relevant resemblance*. The notion of relevant resemblance combines ontological and pragmatic perspectives in modeling: resemblance is an objective matter of fact, while relevance derives from the modeler's interests and goals, purposes and audiences. Among the latter there are goals such as predicting the inflation rate with a degree of accuracy useful for economic policy makers; outlining the core structure of a bubble-generating mechanism in a way that is understandable for the electorate; showing that a result is robust to a change of an assumption so as to impress one's peers; unifying diverse classes of social phenomena so as to expand the

academic authority of economics; designing a regime of regulation for the financial sector useful for legislators.

Relevant resemblance is always incomplete and imperfect. Complete and perfect resemblance is unattainable (and would be impractical anyway), and most partial and imperfect resemblances are irrelevant for a given purpose and audience. A model that relevantly resembles a target, resembles it in a specifically limited way that serves a purpose and helps to reach an intended audience. It highlights only some selected aspects of the real world and does this in some imperfect degree of accuracy. Relevance is a function of the pragmatic context of purposes and audiences, while relevant resemblance is a function of the pragmatic context together with the relationship between the model and the target object. It is the task of model commentary to point out what kind of relevant resemblance is being sought and perhaps achieved (and it is the task of philosophical analysis to investigate whether relevant resemblance can be interpreted as truth or not; for a positive answer, see Mäki 2011a).

This framework also helps to understand the role of false idealizing assumptions in modeling. They are among the descriptions of a model. If we take models to be imagined systems, it is obvious that such systems can be described in many different ways, such as in terms of verbal means, mathematical equations, graphs and diagrams and other visualisations. So idealizing assumptions can be considered as describing or even defining models rather than being statements about some real-world system. The challenge then is one of understanding and justifying such idealizations. The answer lies in the analogy between model and experiment, or what I have called the “experimental moment” in theoretical modeling.

Structurally, theoretical modeling is similar to laboratory experiments. Both aim at the isolation of some important relationship or mechanism – theoretical and material isolation, respectively. Both pursue it by controlling things other than those that are being isolated. Laboratory experimentation does this by *causally manipulating* those “other things” while theoretical modeling does the same by *making assumptions*. There is therefore an obvious sense in which theoretical models are thought experiments. (Mäki 1992)

*Model commentary* is an important part of modeling. It plays a key role in identifying and coordinating the other components of model representation. By specifying the purposes and audiences of the exercise, it fixes the standards of relevance. By illuminating the roles played by idealizing assumptions, it can dispel unnecessary suspicions about some models while helping raise legitimate doubts about others. For example, an informed commentary should be able to show differences between defending an assumption on grounds such as the following four (Musgrave 1981; Hindriks 2006; Mäki 2011b). First, a false idealizing assumption – such as that the information held by economic agents is symmetri-

cal – can be defended by saying that it should be interpreted as the possibly true claim that the asymmetries in information are *negligible* for some given purpose. Second, an assumption can be defended by suggesting that the model is only *applicable* to domains where information is symmetrical or where the asymmetries are negligible. Third, it can be defended by suggesting that, *ceteris paribus*, the use of the assumption makes the modeling of some phenomenon (mathematically) *tractable* (or, in case it additionally distorts non-negligible facts, it should be criticized). Fourth, it can be defended by interpreting it as an *early-step* assumption that will be relaxed in later-step versions of the model; such a de-idealization is a way of de-isolating the model by bringing in previously excluded factors. This may aim at checking the robustness of the model's basic view of the world (see Kuorikoski, Lehtinen and Marchionni 2010) or bringing the model closer to being applicable to some specific domain.

## 4 Economic modeling contested: Three ways

Models are often contested by raising issues of relevant resemblance. A model – or a family of models, or the strategy of modeling – can be challenged by claiming that it fails the test of resemblance relative to some purpose and audience. Or the charge can be that the model commentary has failed to identify the functions of particular idealizing assumptions or the limits of applicability of a model, and so on. Using the account of models and modeling outlined above we can now identify three ways of contesting an economic model, a family of models, or a style of modeling (see Mäki 2009).

The most radical challenge questions the strategy of modeling in general as misguided simply because models and their assumptions are found to be so unrealistic that the strategy is judged to be unsuitable for accessing economic reality. This may manifest the phenomenological worry based of a perceived dissimilarity between the model worlds and the real world, suggesting that *unrealistic models cannot possibly serve as surrogate worlds* that might pave the epistemic way to the real world. To this suspicion my response has been, and continues to be, that unrealistic assumptions per se are no obstacle to successful surrogate modeling.

The second kind of contestation considers unrealistic economic models as potentially successful surrogate objects helpful for acquiring information about the complex real world, but *criticizes particular surrogate models for failure* in the task. While unrealistic assumptions in general are not an impediment to surrogate reasoning about the real world, particular unrealistic assumptions are taken to be the source of failure for they are responsible for the exclusion of important

causal factors from a model. With respect to this charge, it is easy to agree that many models fail just in this way. In the recent years, many very important economic models were blamed for having failed in this manner, reinforcing the practical worry about misguided or missing policy advice.

The third variant of contestation puts forth the charge that *modeling has degenerated into producing and manipulating mere substitute systems* in contrast to surrogate systems. Research and reasoning are concerned with the properties of toy models only, with no further concern about how they relate to and might provide epistemic access to real world systems. Modeling becomes governed by consideration of tractability and mathematical convenience. Again, in response to this worry, it seems obvious that, among the many forces that govern economic modeling, there is also a temptation and tendency in economics to retreat to substitute modeling.

#### **4.1 Can simple models with unrealistic assumptions serve as surrogate systems?**

Simple economic models do not do justice to the rich and complex economic reality. Models involve idealizing assumptions that often severely distort the facts. Models depict closed systems, while the economic world is open and cannot be artificially closed. Therefore, theoretical modeling is a dubious strategy of inquiry in economics. This, or something along these lines, is a critique sometimes put forward against the possibility of successful economic modeling. The focus of this challenge does not lie on the particular idealizing assumptions used in particular models or model families, it rather lies on the strategy itself, that of simplification and idealization in economic modeling.

Consider a model of planetary motion that isolates a simple system that consists of one planet and the sun, both considered as mass points, excluding all other objects and properties and forces other than the gravity between the two included mass points. If the purpose of this model is to provide predictions with a certain degree of accuracy, and if it manages to provide them, it cannot be contested by raising a phenomenological worry just by pointing out that the idealizations of the model distort many features of the actual world.

Then consider the  $2 \times 2 \times 2$  model of international trade. It isolates a system of two countries, two goods and two factors of production (labour and capital), assuming that the factors are homogeneous and production technologies are identical between the two countries and exhibit constant returns to scale. Capital and labour can move within countries but not between them. Competition is perfect within countries, but firms are not considered in the model. There is no unem-

ployment and there are no tariffs. The only difference between the two countries is their relative abundance of labour and capital. This simple Heckscher-Ohlin version of the  $2 \times 2 \times 2$  model isolates a mechanism of comparative advantage that generates outcome patterns in which capital-abundant countries export the products of capital-intensive industries, while labour-abundant countries export the good produced by their labour-intensive industries.

The assumptions of the simple Heckscher-Ohlin model are highly unrealistic and its implied prediction is inaccurate about the actual world. It may be hard to generally justify the idealizations as true negligibility assumptions, claiming that deviations from the facts are generally negligibly small for the predictions of the model to come out sufficiently correct. It may also be hard to find many empirical cases in which the distortions would indeed be negligible, so as to defend them as applicability assumptions. Therefore, they often serve best as early-step assumptions, which are to be relaxed and replaced by other more realistic assumptions. This is what has happened, both in a more piecemeal fashion and in more radical ways, which end up isolating different kinds of mechanism. "New trade theory" relaxes the assumption of constant returns to scale and assumes returns to be increasing. It brings firms to the model, but assumes them to be identical. "New new trade theory" allows for a diversity of firms and analyses their differential roles in relation to international trade. These developments suggest that the models are at least some of the time considered as surrogate systems. Unrealisticness as such does not undermine this ambition.

Perhaps the most striking example of this principle is what has sometimes been called the world's first economic model, J.H. von Thünen's *Der isolierte Staat*, a very simple and highly idealized model of the distribution of agricultural land use (von Thünen 1828; for an analysis, see Mäki 2011a). The model makes highly idealizing assumptions and implies a very idealized land use pattern of concentric rings. Among the idealizations, the region is assumed to be a perfect plain without mountains, valleys or navigable rivers; throughout cultivatable and of homogeneous fertility and climate; to have just one dimensionless town in the middle with a market on which the agricultural products will be sold; to be cut off from the rest of the world by a wilderness. Furthermore, transportation costs and land rents are assumed to be functions of the distance from the town (longer distances are associated with higher transportation costs and lower land rents). And naturally, agents are assumed to be rational maximizers, doing a perfect job in balancing the pull and push of the two magnitudes in deciding where to locate. The assumptions and implications of this simple model are false, but yet there is a fair chance that it manages to isolate a real mechanism that causally contributes to actual land-use patterns. The distortions by the assumptions might not be negligible if the purpose is to predict the outcome pattern with a relatively high

degree of accuracy, but they might be so for the purpose of isolating a fragment of the causal structure of the world.

An important condition for models to succeed as surrogate objects is for the model commentary to be informed about their capacities and limitations, their appropriate domains of application and the sorts of question they can be used to answer. There is a failure of model commentary in case a model is applied to domains to which it does not properly apply and is used for answering explanatory questions on which no illumination can be cast with that model. A good model commentary sees to it that models are applied in a way that promotes the pursuit of the goals for which they are fit.

Theoretical models in economics often provide *how-possibly explanations* – to be pointed out by a commentary. There is an observed pattern, such as a pattern of trade or of agricultural land use. One then suggests a model that depicts a mechanism that has possibly brought about the pattern. No claim is made at this point that the mechanism has actually generated the pattern. Indeed, it might have arisen in some other way as well (such as a land-use pattern having arisen as a result of centralized zoning). Models providing how-possibly explanations are surrogate models for they can be used for making claims about some real structural features of a domain of causes and effects. They often isolate mechanisms but are alone insufficient for determining whether those mechanisms are actually in operation and whether their operation is or is not modified or even overridden by other mechanisms. For these purposes, other models and an informed model commentary are required.

## 4.2 Failing surrogate models with failing unrealistic assumptions

In contrast to the suspicion discussed above, modeling is here not contested as a general strategy that in principle cannot succeed in generating reliable information about the real world. So the possibility of surrogate modeling is granted, but its actual implementation is judged as a failure. All models idealize and are simple, but bad models idealize and simplify too much or in a wrong way. The alleged reasons for such a failure can be many, such as mistaken background theories, incomplete or poor quality data, weak or misguided methods of testing, the tempting mathematical convenience of some idealizing assumptions, ideological bias, and so on.

What many of these criticisms share is the idea that *a model misses some causally important factors that should be modelled*. Another way of putting this is to trace the alleged failure of bad models back to some key assumptions that

are claimed to be responsible for the failure. Those assumptions are idealizations that help exclude from the model world one or more factors that are causally important in the real world.

The flaws of economic models have been diagnosed with respect to the recent crisis in the same way. The two sets of models most often accused of major failure are efficient financial markets models and the macroeconomic DSGE (dynamic stochastic general equilibrium) models. They share the image of unregulated markets as efficient and basically self-correcting, and of economic agents as rational and well informed.

Models of efficient financial markets rely on assumptions such as zero transaction costs and perfect and symmetrical information between the agents. Such idealizations are instrumental in generating an image of the financial system in which market prices fully reflect all available information and in which there can be no bubbles in asset prices such as those of stocks or houses. This is a surrogate object that has the nice feature of being self-regulated and having the capacity of containing all relevant risks.

It then takes a major step to move from this surrogate world to the real world. This step can be taken in a variety of ways and on a number of grounds. One extreme and straightforward option would be to reject the model on phenomenological grounds, simply because the key idealizing assumptions seem to get the facts wrong. At the other extreme, without much further investigation, the model would be accepted as a true or useful surrogate system that is relevantly similar to the relevant target systems. It would be believed to get the important properties of the real financial system right – such as asset prices reflecting all available information, no bubbles being generated, and so the real-world financial system having the self-stabilizing properties needed for containing all risks.

The critics claim that economists or practitioners in the financial markets – enchanted by models of efficient markets – have been too hasty in concluding that there is relevant resemblance between the models and the real world, perhaps believing that informational imperfections in real-world markets are negligible. The critics believe there is no relevant resemblance at all, so the real-world imperfections are far from negligible. They argue that the properties of real-world markets may in fact be the reverse of those of the model-world markets:

... where the Efficient Markets Hypothesis suggests that financial markets provide a way of managing economic risk, the evidence suggests that they are actually a major source of risk. (Quiggin 2010, 51)

The charge might be put by saying that *economists have missed real-world risks by underestimating modeling risks*. The move from the model world to the real world



is typically far more difficult and risky than is the mere production of publishable results of the theoretical examination of models. It is these epistemic risks that may have been neglected.

The same complaint can be made about dominant macroeconomic models. In an interview in 2009 Nobel Laureate Robert Solow diagnoses their failures as deriving from their shared image of the economy that distort some basic facts:

currently fashionable macroeconomics likes to formulate things in a way that inevitably endows the economy with more coherence and purpose than we have any right to assume.

By saying that, “without any right” the models “endow the economy” with properties that the economy does not have, Solow implies that macroeconomists have been careless risk takers in moving from examining their well behaving model worlds – in which there is a lot of “coherence and purpose” – to making claims about the less orderly real world.

The contested macroeconomic models rely on the image of financial markets being efficient, so no further inquiry is required to incorporate more nuanced assumptions about how the financial markets actually function. This is not the only objection. It is an instance of the more general complaint that the models leave out causally important factors and in doing so also miss important explananda. Those factors are causally important, because they are responsible for, say, the sort of crisis we have recently witnessed. Since the causes of such crises are not among the isolated factors in the models, their effects – the crises and their characteristics – cannot be explained or predicted. In the worst case, they cannot even be conceived within the framework of those models. This is, among many others, the main focus of the complaints levelled by Nobel Laureate Joseph Stiglitz and many others.

Macroeconomic models using representative agents miss the crucial causal factors that lie in things such as informational asymmetries, structure of financial markets, and corporate governance. These models therefore do not recognize phenomena such as excess indebtedness, debt restructuring, bankruptcy, and agency problems. Any model with these characteristics

leaves out much, if not most, of what is to be explained; if that model were correct, the phenomena – the major recessions, depressions and crises that we seek to understand – would not and could not have occurred (Stiglitz 2011, 168).

These models fail to incorporate factors that are crucial for major macroeconomic fluctuations and instead focus on minor price distortions due to inflation. Macroeconomic models are better in explaining “the small and relatively unimportant fluctuations that occur ‘normally’, ignoring the large fluctuations that have

episodically afflicted countries all over the world.” Those models have failed, and are unable, to answer explanatory questions such as, Why have such fluctuations occurred? Why do disturbances get amplified? And why are recoveries so slow? (Ibid., 169.)

So the core of the contestation is one of failed isolation: the poor models have isolated factors of secondary importance and by idealizing wrongly have come to leave out many others that are crucial. The charge is not that models fail to represent or are not intended as surrogate objects but rather that “the conventional models inadequately modelled – and typically left out – many, if not most, of the key factors that played a central role in this crisis” (172). The issue is about the relevant resemblance between the models and the target phenomena, and the claim is that the issue has been unsuccessfully resolved. Given that relevance is determined by the explanatory urge to understand the behaviour of the bubbles of the current crisis, the verdict levelled by the critics is that for this purpose, models do not relevantly resemble their targets.

The reason why a model fails is that the causal factors it excludes are not negligible. This is what Stiglitz implies:

Economists assumed that information was perfect even though they understood that it was not. Theorists hoped that a world with imperfect information was very much like a world with perfect information – at least so long as the information imperfections were not too large. (2010, 242)

The charge is here that economists dealt with the false perfect information assumption as a true negligibility assumption. But as Stiglitz reminds us, economists have no rigorous way of measuring the size of information imperfections – which makes estimating their negligibility even more difficult. This creates room for the role of sheer hope that they are negligible (yet Stiglitz himself does not hesitate to claim that information imperfection is not negligibly small).

The issue often becomes transformed into an issue of the purposes of modeling and the intended domain of their applicability.

Is the purpose of an economic model to help us predict a little bit better how the economy is performing in ‘normal’ times – when things do not matter much? Or, is the purpose of an economic model to predict, prevent and manage big fluctuations and crises? (Stiglitz 2011, 168)

The criticism is often phrased by saying that the poor models deal with “special cases where market inefficiencies do not arise” (Stiglitz 2011, 166) or that they do not apply to economies that are capable of generating bubbles. In the imagined worlds of these models, agents are super-rational and fully informed, there are

markets for all goods and all risks extending infinitely far into the future and covering all risks (“one can buy insurance against every conceivable risk”). In such worlds, bubbles don’t occur (Stiglitz 2010, 252). Careless epistemic risk taker economists then proceed to conclude that bubbles do not occur in real world economies, either.

Such careless risk taking reflects deficiencies in the model commentary that fails to inform modelers and model users about the structure of the modeling exercise and what it takes to successfully manage the epistemic risks in model application. Among other things, the commentary should give the obvious advice to build a pool of models from which one can choose and put in use those that are appropriate to the kind of case at hand – for example, a set of models for situations with bubble-generating mechanisms in operation and another set for other sorts of situation (cf. Colander 2010). This advice may fail to be given insofar as it cannot be easily reconciled with disciplinary conventions such as that of unification. Some observers suspect that behind such an uninformed model commentary there is an ideological bias: “Unfortunately, careful attention to the limitations of simplified models has not been the norm in the era of market liberalism.” (Quiggin 2010, 109)

### 4.3 Substitute modeling

The type of challenge discussed in the previous section identifies possible modeling failure in the failed attempt to build models that would isolate the factors that are causally important for some major phenomena such as financial crises of the present type. While such a failure is a matter of a *failed attempt*, the one to be briefly discussed in this section is a matter of *failing to attempt*. This is the distinction between surrogate modeling and substitute modeling (Mäki 2009).

*Surrogate modeling* is motivated by epistemic ambitions that reach beyond learning about just the model. By directly examining the properties of the model, the modeler seeks to indirectly learn about some target. Resemblance between the model world and the real world is an issue that is prompted and perhaps settled. In surrogate modeling, the model system is intended – or found to serve – as a *bridge* to some real system (and may fail as such a bridge).

By contrast, *substitute modeling* is a degenerate activity that has no ambitions beyond dealing just with models. The modeler only examines the model and only learns about its properties, whereas the resemblance of the model world with some real world system is not prompted as a genuine issue to be resolved. Examining a model is a substitute for trying to indirectly access the real target. Criteria other than those indicating resemblance dominate the exercise. Rather

than offering a bridge, the model remains an intellectual *island* unconnected to the real world.

There is a deeply rooted suspicion among many critics that much – or at any rate too much – economic modeling is of a substitute variety. According to this charge, economists too often only have an interest in examining the properties of their models and have no interest in checking how those properties relate to the properties of some important real world systems. At the time of crises, this charge regularly makes an appearance (e.g. Hodgson 2009).

This provides a framework for reading Nobel Laureate Paul Krugman's critical account of the sources of failure of economics in dealing with – anticipating and analyzing – the financial and economic crisis of 2007–2008. In a column in the *New York Times Magazine*, Krugman (2009) stated that

[...] the economics profession went astray because economists, as a group, mistook beauty, clad in impressive-looking mathematics, for truth. [...]

This can be translated into the idea that economists have failed in dealing with the crisis because they have been busy with substitute modeling rather than surrogate modeling. Accordingly, economists have been preoccupied with the beauty and neatness of their models, expressed in impressive mathematics, while this has contributed nothing to the task of finding relevant truths about the real world.

Regarding the contents of these models, Krugman says economists have envisaged a fantasy world of perfectly rational agents in perfectly functioning markets, very far removed from the imperfections of the real world – and that this must change.

When it comes to the all-too-human problem of recessions and depressions, economists need to abandon the neat but wrong solution of assuming that everyone is rational and markets work perfectly. The vision that emerges as the profession rethinks its foundations may not be all that clear; it certainly won't be neat; but we can hope that it will have the virtue of being at least partly right.

So the model worlds envisaged by economists – with perfect rationality and perfect markets, and therefore without the sorts of financial bubble that burst in 2008 – have been excessively neat and tractable. Such models permit relatively easy derivations of relatively unambiguous modeling results. Krugman might be taken to suggest that there is some sort of trade-off between neatness and truth, such that when trying to get their models closer to the truth (“at least partly right”) economists will have to give up at least some of the neatness of their models.

We may develop this line of thought further by envisaging an extreme situation in which the virtues of neatness and tractability completely come to dominate modeling at the expense of other (“reality-oriented”) virtues. Once a model world is sufficiently far from the real world, the modeler is tempted to pay all her attention to the properties of the models only and to ignore any further issues of resemblance with the real world. This would be degenerate substitute modeling.

This inclination could be generated or reinforced by an *excessive role of mathematical convenience or tractability* in modeling. Some idealizing assumptions in models are indeed made to serve mathematical tractability purposes (they are called “modelling tricks” by Krugman; cf. Mäki 1992; Hindriks 2006). In case tractability and negligibility do not coincide – in case the distortions brought about by those tractability idealizations are not negligible from the resemblance point of view – we have a possible source of failure (Mäki 2011b). John von Neumann – surely with no dislike for mathematics per se – was aware of these dangers:

As a mathematical discipline travels far from its empirical source, or still more, if it is a second and third generation only indirectly inspired by ideas coming from “reality”, it is beset with very grave dangers. It becomes more and more purely aestheticizing, more and more purely *l’art pour l’art*. This need not be bad if the field is surrounded by correlated subjects, which still have closer empirical connections, or if the discipline is under the influence of men with an exceptionally well-developed taste. But there is a grave danger that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganized mass of details and complexities. In other words, at a great distance from its empirical source, or after much “abstract” inbreeding, a mathematical subject is in danger of degeneration. (von Neumann 1947, 9)

Yet, we should not rush to any simplistic conclusions on this matter. The world of modeling – not just the world modelled – is complex and easily misunderstood. It is fairly safe to make the general observation that economists are happy with examining models and making claims about their properties in a rigorous manner, but are equally happy with saying nothing – or at most saying something very casual – about any real targets based on what they discover about models (cf. Sugden 2009). Yet, as such, this alone does not imply that economists are practicing substitute modeling. Let me explain why.

Talking about models as if they were the world is a natural aspect of model-based research strategy in all disciplines. Models easily become objectified or reified as the immediate targets of inquiry: their properties and behaviour are investigated and the results are reported in scientific publications. The important question is *what else* is going on in inquiry. The relevant dimensions of the possible answers to this question are the collective and the historical. There is the

collective dimension: *What does the research community do as a whole?* And there is the historical dimension: *What will happen in later stages of research?*

Some portion of economic modeling could perhaps be saved from charges of substitute modeling provided one or both of the following two conditions were met: there is a well functioning division of intellectual labour such that while some economists only build and examine models, there are others doing the hard work of investigating how those models relate to the real world; and/or there is a historical sequence of bodies of research such that an earlier stage of study of model properties will in due time be followed up by the study of how those model properties relate to real world properties. Another way of putting this is to say that substitute modeling may only appear to be such, while in fact it is a phase of surrogate modeling considered in a broad enough collective and historical context.

Indeed, in their defensive commentary of apparent substitute models, economists often appeal to such collective and historical considerations. However, this is an all too easy move if nothing more specific is said about the two dimensions. The critic may grant the relevance of the collective division and historical ordering of tasks, and yet argue that economic modeling has recently failed just in this. The needs of policy are often urgent, so they cannot wait for some possible future generation of economists to do its share in bridging the gap between the models and the world. This would be a failure in the institutions of modeling – its rules and conventions, incentive structures and industrial organization.

## 5 Conclusion

I have outlined three sets of ideas. First, I have articulated what I think are among the dominant disciplinary conventions that guide economic modeling. Second, I have sketched a general account of modeling as ontologically and pragmatically constrained epistemic activity. Third, without trying to be exhaustive, I have provided a rough partial typology of three ways of contesting economic modeling: questioning the use of unrealistic assumptions and thereby the strategy of modeling as such; questioning the use of particular unrealistic assumptions and models conceived as surrogate objects; and questioning the allegedly degenerate practice of substitute modeling.

The boundaries between the three ways of contestation are not always clear and sharp. For example, it is not always easy to tell a surrogate model used for offering a how-possibly account from a substitute model governed by goals other than reasoned truth about the world. More generally, the difference between the two may be hard to tell, because the collective and historical dimensions of

excuse allow for flexible interpretations: there is no unambiguous and uncontroversial way of fixing the required sort of division of research labour and the permitted time lag between examining a model and checking how it relates to the real world. Yet I believe something of the sort I have suggested might serve as a beginning for drawing a map within which various ways of contesting modeling might find a place.

As I see it, modeling is a powerful and indispensable method of *managing complexity* in a discipline like economics. At the same time, it is extremely important to recognize the difficulties of *managing the risks of modeling*. As Keynes said, in the long run we are all dead. The critic of modeling might add that in the long enough run, we may all be killed by some deep economic disaster – the possibility which economists failed to conceive and the actual occurrence which they failed to anticipate just because they were too fond of their simplistic model worlds.

## Acknowledgement

Work on this paper has been sponsored by the Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences.

## References

- Akerlof, G. & Shiller, R. (2009). *Animal Spirits. How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*. Princeton: Princeton University Press.
- Colander, D. (2010). The economics profession, the financial crisis, and method. *Journal of Economic Methodology* 17(4). 419–427.
- Farmer, J. Doyne & Fole, Duncan (2009). The economy needs agent-based modelling, *Nature* 460 (August). 685–686.
- Giere, R. (1999). *Science Without Laws*. Chicago: University of Chicago Press.
- Hindriks, F. (2006). Tractability assumptions and the Musgrave-Mäki typology. *Journal of Economic Methodology* 13. 401–423.
- Hodgson, G. M. (2009). The great crash of 2008 and the reform of economics. *Cambridge Journal of Economics* 33. 1205–1221.
- Krugman, P. (2009). How did economists get it so wrong? *The New York Times Magazine*. 6 September.
- Knuuttila, T. (2009). Representation, idealization, and fiction in economics: From the assumptions issue to the epistemology of modeling. In: Suárez, M. (ed.): *Fictions in Science. Philosophical Essays on Modeling and Idealization*. London: Routledge. 205–231.

- Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *British Journal for the Philosophy of Science* 61. 541–567.
- Lehtinen, A. & Kuorikoski, J. (2007). Computing the perfect model: Why do economists shun simulation? *Philosophy of Science* 74. 304–329.
- Mäki, U. (1992). On the method of isolation in economics. *Poznan Studies in the Philosophy of the Sciences and the Humanities* 26. 319–354.
- Mäki, U. (2005a). Reglobalising realism by going local, or (how) should our formulations of scientific realism be informed about the sciences. *Erkenntnis* 63. 231–251.
- Mäki, U. (2005b). Models are experiments, experiments are models. *Journal of Economic Methodology* 12. 303–315.
- Mäki, U. (2009). MISSING the world: Models as isolations and credible surrogate systems. *Erkenntnis* 70. 29–43.
- Mäki, U. (2011a). Models and the locus of their truth. *Synthese* 180. 47–63.
- Mäki, U. (2011b). The truth of false idealizations in modelling. In: Humphreys, P. & Imbert, C. (eds.). *Models, Simulations, and Representation*. London: Routledge. 216–233.
- Morgan, M. S. (2003). Experiments without material intervention: model experiments, virtual experiments and virtually experiments. In: Radder, H. (ed.). *The philosophy of scientific experimentation*. Pittsburgh: University of Pittsburgh Press. 216–235.
- Morgan, M. S. & Knuuttila, T. (2011). Models and modelling in economics. In: Mäki, U. (ed.). *Handbook of the Philosophy of Economics*. Elsevier. 49–88.
- Musgrave, A. (1981). 'Unreal assumptions' in economic theory: The F-twist untwisted. *Kyklos* 34. 377–387.
- Neumann, J. von (1947/1961). The Mathematician In: Neumann, J. von. *Collected Works*. Vol. I. Edited by Abraham H. Taub. Oxford: Pergamon Press. 1–9.
- Plessis, S. du (2010). Implications for models in monetary policy. *Journal of Economic Methodology* 17(4). 429–444.
- Quiggin, J. (2010). *Zombie Economics. How Dead Ideas Still Walk Among Us*. Princeton: Princeton University Press.
- Stiglitz, J. E. (2010). *Freefall. Free Markets and the Global Economy*. London; etc.: Penguin Books.
- Stiglitz, J. E. (2011). Rethinking macroeconomics: What went wrong and how to fix it. *Global Policy* 2. 165–175.
- Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis* 70. 3–27.
- Thünen, J. H. von (1910) *Der isolierte Staat in Beziehung auf Landwirtschaft und National-ökonomie*. Jena: Verlag von Gustav Fischer.

### **Prof. Dr. Uskali Mäki**

Centre of Excellence in the Philosophy of the Social Sciences (TINT)

POB 24

00014 University of Helsinki

Finland

uskali.maki@helsinki.fi



Julian Reiss

# Models, Representation, and Economic Practice

## Commentary on Uskali Mäki

Few, if any philosophers of economics and economic methodologists have been brought up without being nurtured on Uskali Mäki's writings on idealisation, models, realism, truth, isolation and many other aspects of economic methodology. I certainly have been. In graduate school, his article 'Scientific Realism and Some Peculiarities of Economics' (Mäki 1996) was presented to me as a classic, and I still use it to teach my own students about realism. It is therefore a particular pleasure to have been given the opportunity to provide some thoughts on Mäki's latest on models and idealisation.

The aim of Mäki's paper is three-fold. First, he outlines a number of epistemic virtues economists seek in models – such as being constrained by theory, being parsimonious, broadly applicable, couched in mathematics, uninfluenced by findings in other disciplines as well as providing insights into phenomena and their generative mechanisms – as well as obstacles to their realisation, e.g. the Duhem-Quine problem. Second, he gives a new formulation of his account of modeling and defends some of its aspects. Finally, he discusses three challenges critics have posed to economic modellers and either rebuts or sustains these challenges. In my comment I will focus on the account of modeling and how it deals with the three challenges.

## 1 Representation by Models

For convenience let me repeat Mäki's account of representation here:

### [ModRep]

Agent *A* uses (imagined) object *M* as a **representative** of (actual or possible) target *R* for **purpose** *P*, addressing **audience** *E*, at least potentially prompting genuine **issues of resemblance** between *M* and *R* to arise, describing *M* and drawing inferences about *M* and *R* in terms of one or more **model descriptions** *D*, and applies **commentary** *C* to identify and coordinate the other components.

[ModRep] departs, more than any other account, from the usual way of thinking of representation as a two-place relationship between a model and a target. Why two additional places should be included is straightforward. Nothing is a model of something else without being stipulated as such – or ‘used as a representative’ as Mäki states. Thus, agency is essential. And there is no way to tell whether a model is a good one, or whether it is an accurate or adequate representation, without specifying a purpose.

The reasons for including the other aspects are less conspicuous, despite what Mäki remarks on pp. 91–94. One way to understand them would be to hold that representation is not a four-place relationship between a model, a target, a user or agent, and a purpose, but rather a seven-place relationship that also includes an audience, a description and a commentary.

If so, we may ask why we need an audience in addition to a purpose. I was taught about the solar system using a mainly mechanical model that had a big light bulb at its centre to represent the sun. I do not think that model was of much scientific use. Its purpose was to help the teacher getting across some basic astronomical knowledge to the students. The purpose, properly specified, includes the audience. We could ask: does the representation relation change when the audience changes? Surely not when the solar system model is used between 9 a.m. and 10 a.m. for one set of students and between 10 a.m. and 11 a.m. for another. But it would make a difference if a crazy scientist, who would use the model to draw inferences about how to send a probe to Mars, replaced the students. However this would be because the model was built for the purpose of education (or, more specifically, of educating grade-eight grammar school students in the UK with such-and-such a background and ...) and not for calculating the trajectory of Mars probes. Thus, purpose includes audience.

The same is true for the commentary. A commentary can draw our attention to the fact that a model was built for one purpose and not for another. David Colander suggested it might be a good idea, if an economic model included ‘warning labels to prevent the model from being misused’ (Colander 2010). The fact that a model was built for one purpose and not another makes some applications of the model instances of misuse, not the commentary. A different commentary does not change the representation relation as long as the purpose stays the same.

Finally, it is true that models always occur under a certain description. But it is noteworthy that the descriptions define the model. It is not the case that the same model *M* has a different representation relation to its target *R* when the description of it is changed. Rather, the model has changed – and it is *qua* that change that the representation relation may or may not have changed.

Therefore, Mäki cannot mean that representation is a seven-place relationship. Instead, I would suggest, the somewhat Baroque account is meant to remind

us that representation is a complex scientific activity that cannot be reduced to the simply minded search for similarity relations between one object (the ‘model’) and another (the ‘target’). This, to my mind, is entirely correct and an important point to make.

The representation relation itself has, according to Mäki, two aspects: the ‘representative’ aspect and the aspect of ‘resemblance’. The agent decides whether an object is a representative of another. This is why I stated above that agency was essential. At this stage there are not yet limits to what can be used as a model of something else. If I point to some sprawling weeds in my garden and say: ‘These weeds are a model of world capitalism!’, then the weeds *are* a model of world capitalism. *Whether* something is a model of something else is decided by fiat or use. If someone countered my exclamation with: ‘No, they are not!’, he would not have understood the rules of the game. But of course, that something is a model of something else does not automatically mean that it is also a *good* model of it. This is why in addition to representativeness we need resemblance.

Whether or not any given model is also a good model of its target depends in part on the purpose of its use. If my intention is to suggest that capitalism takes over even the remotest corners of the world economy, like the weeds are taking over my garden, then they might well be a good model. But certainly I will not learn many useful things about the causes of world capitalism’s behaviour by examining my weeds.

Now, while Mäki does not address this issue explicitly, his paper suggests that he takes the circumstance whether the model resembles its target to be a fact about the relation between the model and its target. Pragmatic factors determine, which aspects of the model (or the model/target relationship) are relevant. But once this issue is settled, facts alone determine whether a model bears ‘resemblance’ to its target, and thereby whether the model is a good one.

In my view, this gives context and purposes a too small role to play in representation. Resemblance is not a natural kind whose presence is determined by the facts alone. Any two objects are similar and dissimilar in uncountable ways. We need context and purpose to determine not only what aspects are relevant, but also in what sense model and target should resemble each other. Paul Teller makes the point succinctly as follows (Teller 2001: 402):

In short, once the relevant context has been specified, for example by saying what is to be explained or predicted and how much damage will result from what kinds of error, the needs of the case will provide the required basis for determining what kind of similarity is correctly demanded for the case at hand. More specifically, similarity involves both agreement and difference of properties, and only the needs of the case at hand will determine whether the agreement is sufficient and the differences tolerable in view of those needs. There can be no general account of similarity, but there is also no need for a general account

because the details of any case will provide the information, which will establish just what should count as relevant similarity in that case. There is no general problem of similarity, just many specific problems, and no general reason why any of the specific problems need be intractable.

## 2 Challenging Economic Models

Mäki discusses three sets of criticisms commentators have levelled against economic models. The first challenge is that economic models cannot be epistemically useful, because they simplify and idealise. The second holds that the particular assumptions a particular model makes may be unsuitable for a particular task at hand. The third criticism concerns the economics profession at large and maintains that economists practise modeling too much (or even exclusively) for its own sake rather than with specific (policy or other practical) applications in mind. Let us consider these in turn.

(A) Simple and idealising models (SIMs) cannot be epistemically useful.

To begin with a disclaimer, I do not know anyone who makes the criticism at this very general level, and Mäki does not provide a single reference to anyone who does. The charge is highly implausible: all models simplify and idealise in myriad ways (for a classification, see for instance Wimsatt 2007: 101–102), and it would be hard to maintain that no model is epistemically useful. Perhaps the charge is somewhat more specific: all *economic* models simplify and idealise too much *relative to economic reality*. Again, I know no one who would hold such an implausible view. Tony Lawson (1997, 2003) comes close, but his criticism concerns the mathematisation of economic models, not models (or simplification/idealisation) *per se*. All other critics I am aware of make more nuanced remarks, remarks concerning specific modeling strategies and specific domains of application.

Mäki nevertheless provides two defences. First, even the most highly simplifying and idealising assumptions may be considered to be mere ‘early step assumptions’, which are to be replaced by more realistic assumptions at a later stage. Second, SIMs often provide ‘how-possible explanations’.

Neither line of defence is entirely convincing. The ‘SIMs play a heuristic role for future models that are epistemically useful’-defence leads into a dilemma. Either simplifications and idealisations can be relaxed so as to make models more realistic and thereby epistemically useful or they cannot. If they can, one would

have to be able to tell a good story of what precise heuristic role the SIMs play on the road to better models (why do we build epistemically useless models if we can have useful models?), and I would imagine this will be no mere trifle. More importantly, however, one could simply ignore SIMs and focus on those models that are useful. The criticism would amount to saying no more than ‘some models are not useful’. But that’s hardly a criticism.

Or the simplifications and idealisations cannot be relaxed. But then models involving them could not play a heuristic role for better models. Either way, this line of defence leads straight into a *cul-de-sac*.

The other defence is that SIMs can provide ‘how-possible explanations’. But this notion has the modal operator in the wrong place. A ‘how-possible “explanation”’ is not an explanation. It is *possibly* an explanation. Suppose an implication of a SIM is a claim of the following form: ‘In situation *S* (which can be described by conditions *s*<sub>1</sub>, *s*<sub>2</sub>, ..., *s*<sub>*n*</sub>), factor *C* causes outcome *E*’. This allows us to provide a more precise characterisation of what a SIM is: a model, which entails causal claims that are true under conditions rarely or, more frequently, never found empirically. Typical examples of such conditions include a continuum of economic agents, agents who are perfectly rational, agents with an infinite lifespan, businesses located on a line that has neither depth nor breadth, consumer goods that have a single property and so on.

There is indeed a sense in which SIMs make a possibility claim, viz. they show that *it is possible that C causes E*. It is important to see, however, that this is a very weak sense of possibility, something like logical or conceptual possibility. SIMs do not prove an existence claim of the form: ‘There is an empirical situation *S*<sup>c</sup> in which *C* causes *E*’.

I have given an account of how models, resulting in those possibilities, can be epistemically useful (Reiss 2008: Ch. 6). Essentially, if everyone in some epistemic community at some point in time is convinced that it is impossible that *C* causes *E*, it might well be useful to learn that there are conditions, even though non-empirical conditions, under which *C* does cause *E* because now we have a reason to investigate empirically whether *C* causes *E* in situations that interest us. A SIM, in my 2008 terminology, gives *prima facie* evidence for a causal claim: a licence to further investigate it. (Till Grüne-Yanoff 2009 gives a very similar albeit more detailed account.)

But this account is not Mäki’s. To show that a causal relation is logically or conceptually possible is not to explain anything. Take the famous Akerlof lemons model in which asymmetric information brings about market failure (Akerlof 1970). Mäki is of course right to say that Akerlof provides an account of how market failure might come about. But that account explains no single instance of market failure. Rather, it gives a possible or *potential* explanation. A potential

explanation is not a genuine explanation unless all other potential explanations have been ruled out. Therefore, the ‘how-possible explanation’ defence does not work, either.

- (B) Specific SIMs simplify and/or idealise too much or in the wrong way (to be useful for the task at hand)

For this charge to have any bite, one must couple it with the empirical claim that such ‘bad’ models are typically used in epistemically or practically important applications or both. One does not have to go far afield to find some evidence for that empirical claim in the current situation in which blaming economists and their modeling practices for the financial crisis of the late 2000s has become an academic fashion (see for instance Acemoglu 2011; Akerlof and Shiller 2009; Cassidy 2009; Colander 2010; Colander et al. 2009; du Plessis 2010; Hodgson 2011; Kirman 2010; Lawson 2009, Ormerod 2010; Roubini and Mihm 2010; Stiglitz 2009, 2010, 2011). Mäki joins this choir, but goes beyond many of the other commentators for he provides a general methodological account why it is the case ‘that macro and financial economists helped cause the crisis, that they failed to spot it, and that they have no idea how to fix it’ (*The Economist*, July 16 2009): their models exclude non-negligible causal factors.

Here Mäki’s realism stands in the way of a more nuanced analysis. Economists are not in the business of building models that represent all and only those factors that are causally relevant for outcomes of interest. They are in the business of building models that describe and predict, explain and underwrite policies. Non-negligible causal factors will no doubt play *some* role in such models. But, depending on the purpose, such factors will often play an attenuated or negligible role.

We all know that one does not need causality for predictive success. For a classical philosophers’ example, the barometer reading reliably predicts the storm without causing it. If the goal is to predict a storm, there is little reason to model all the causally relevant factors for storm. It is indicator variables we need, and barometers are good indicators.

The problem for the causal realist is that factors that *cause* outcomes of interest are never essential and often of limited usefulness. Explaining phenomena is the best test case, because of the tight semantic connection between ‘causes’ and ‘explains’. But even for explanation causation is not essential. Though highly successful, the causal account of explanation is not the only existing one. Most famously, there is the alternative account of explanation as unification (Friedman 1974; Kitcher 1985; 1989). This is, of course, not an argument in itself, as the unification account might just be wrong. But what is important to understand is,

that the causal account is difficult to square with the fact that all models idealise heavily and yet appear to be explanatory – and are often taken to be explanatory by economists and many other scientists (see Reiss 2012 for a discussion; for a defence of the causal account in the light of idealisations, see Strevens 2007). The causal account is also difficult to square with the fact that some relations seem to require non-causal, but explanatory relations such as constitution (Ylikoski 2011). The least we should take from these considerations is that not all successful explanations are causal.

Finally, it is clear that successful *descriptions* do not always require causal information (for an argument to the effect that causal-mechanistic information is not always helpful for description, see Reiss 2007), and it has been argued that causal relations are not needed for policy analysis (Leuridan et al. 2008).

The upshot of this discussion is that a more nuanced argument is needed to support sweeping claims of the sort ‘economists’ models helped cause the crisis’. There is no unique recipe for failure one might say. Omitting non-negligible causally relevant factors in a model may well be a reason for failure. Only in the light of a specific use of the model and if an argument to the effect that the omission was essential for the failure is available, we can determine whether this is so.

### (C) Modeling for modeling’s sake

The final challenge is normative. A model, as we have seen, is good or bad only in the light of the purpose pursued with its construction and use. Any related methodological criticism is consequently instrumental: we criticise models not as such but rather as means to given ends. But many methodological debates concern in fact the ends themselves: do we want models to describe and predict or shall we seek explanation in addition (famously, Friedman 1953)? Shall we, perhaps, ultimately seek only explanation, because prediction is impossible and description is subsidiary (e.g., critical realism: Lawson 1997; 2003; the new mechanists: Elster 2006; prediction is impossible: McCloskey 1998)? *Is* accurate description of merely instrumental value or is it an important end in itself (e.g., Sen 1981)?

The present charge concerning the aims of economic modeling is that economists too often engage in the pursuit of models with primarily non-empirical virtues at the expense of the more empirical description/prediction/explanation/policy analysis. What these non-empirical virtues are is not quite clear. In a widely cited op-ed piece, Paul Krugman lamented that economists were ‘mistaking beauty for truth’ (Krugman 2009; see also Juselius 2009). Mäki puts it differently: economists too often pursue *substitute* in lieu of *surrogate* modeling. A model is usually a model for or of something. An animal model is called such because it is examined in order to make predictions about other animals,

usually humans. The mechanical model with the lamp at its centre I mentioned above was a model of the solar system. Mäki calls this kind of modeling practice in which a model is used as a stand-in for something else, ‘surrogate modeling’. In surrogate modeling, we examine one system to learn about another, because the latter is epistemically inaccessible for technological, financial or ethical reasons.

‘Substitute modeling’ works without target systems of interest. The model system’s properties are examined, but not for the sake of learning about another system. The model system’s properties are examined for the sake of learning about the model system.

Mäki calls this activity ‘degenerate’ (p. 101). This is a little too fast, however, according to Mäki’s considerations that follow. Perhaps there is some sort of division of labour going on between ‘theoretical’ economists devising models and investigating their properties and ‘applied’ economists using the models to describe, predict and explain phenomena of interest, and to prepare policy.

Something goes wrong only when the discipline as a whole becomes one of substitute modellers – because no one is left to apply the models to our urgent practical and policy problems (p. 104).

Though I share Mäki’s concern in principle, I would like to add that divisions of intellectual labour of the proposed kind often only happen to the detriment of the practical goals of a science (Kitcher 2001; Cartwright 2006; Reiss 2008: Ch. 5). Simply because of the way in which science proceeds, one cannot easily separate the more theoretical role of constructing problem-solving templates of wide applicability and the more applied role of using these templates for solving concrete problems. New models always build on old models; and if all the models there are were built with a particular purpose in mind, it is very difficult to build new models for different purposes (Biddle and Winsberg 2009). One would have to start from scratch. But letting applied scientists build models that are useful to them from scratch is exactly what the proposed division of labour is meant to prevent.

Robert Sugden senses this problem when he writes (Sugden 2009: 25):

In the light of Schelling’s argument about social mechanisms, however, I cannot claim that theorists who make such claims are necessarily committing methodological errors or failing to act in good faith. It is just that the approach of looking for significant mechanisms while not trying to explain anything in particular seems unlikely to be productive.

My point is stronger: to build models with no particular application in mind is to commit a methodological error – as long as the aims of economics are considered to be largely practical.



## References

- Acemoglu, D. (2011). The Crisis of 2008: Lessons for and from Economics. In: Friedman, J. (ed.). *What Caused the Financial Crisis?* Philadelphia: University of Pennsylvania Press. 251–261.
- Akerlof, G. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84. 488–500.
- Akerlof, G. & Shiller, R. (2009). *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*. Princeton: Princeton University Press.
- Biddle, J. & Winsberg, E. (2009). Value Judgments and the Estimation of Uncertainty in Climate Modeling. In: Magnus, P. D. & Busch, J. (eds.). *New Waves in the Philosophy of Science*. New York (NY): Palgrave MacMillan.
- Cartwright, N. (2006). Well-Ordered Science: Evidence for Use. *Philosophy of Science* 73. 981–990.
- Cassidy, J. (2009). *How Markets Fail: The Logic of Economic Calamities*. New York (NY): Farrar, Straus and Giroux.
- Colander, D. (2010). The economics profession, the financial crisis, and method. *Journal of Economic Methodology* 17(4). 419–427.
- Colander, D., Goldberg, M., Haas, A., Juselius, K., Kirman, A., Lux, T., & Sloth, B. (2009). The Financial Crisis and the Systemic Failure of the Economics Profession. *Critical Review* 21(2–3). 249–267.
- Elster, J. (2006). *Explaining Social Behaviour*. Cambridge: Cambridge University Press.
- Friedman, M. (1953). The Methodology of Positive Economics. In: *Essays in Positive Economics*. Chicago, IL: Chicago University Press. 3–43.
- Grüne-Yanoff, T. (2009). Learning From Minimal Economic Models. *Erkenntnis* 70. 81–99.
- Hodgson, G. (2011). Reforming Economics after the Financial Crisis. *Global Policy* 2(2). 190–195.
- Juselius, K. (2009). Is Beauty Mistaken For Truth? A Marchallian [sic] Versus a Walrasian Approach to Economics. Available from: <http://causesofthecrisis.blogspot.com/2009/10/katarina-juselius-on-is-beauty-mistaken.html>.
- Kirman, A. (2010). The Economic Crisis is a Crisis for Economic Theory. *CESifo Economic Studies* 56(4). 498–535.
- Kitcher, P. (2001). *Science, Truth and Democracy*. Oxford: Oxford University Press.
- Lawson, T. (1997). *Economics and Reality*. London: Routledge.
- Lawson, T. (2003) *Reorienting Economics*. London: Routledge.
- Lawson, T. (2009). The current economic crisis: its nature and the course of academic economics. *Cambridge Journal of Economics* 33(4). 759–777.
- Leuridan, B., Weber, E., & Van Dyck, M. (2008). The Practical Value of Spurious Correlations: Selective versus Manipulative Policy. *Analysis* 68. 298–303.
- Mäki, U. (1996). Scientific realism and some peculiarities of economics. In: Cohen, R. S. et al. (ed.) *Realism and Anti-Realism in the Philosophy of Science. Boston Studies in the Philosophy of Science* 169. Dordrecht: Kluwer. 425–445.
- McCloskey, D. (1998). *The Rhetoric of Economics*. Madison, WI: University of Wisconsin Press.
- Ormerod, P. (2010). The Current Crisis and the Culpability of Macroeconomic Theory. *Contemporary Social Science: Journal of the Academy of Social Sciences* 5(1). 5–18.
- Plessis, S. du (2010). Implications for models in monetary policy. *Journal of Economic Methodology* 17 (4). 429–444.

- Reiss, J. (2007). Do We Need Mechanisms in the Social Sciences? *Philosophy of the Social Sciences* 37(2). 163–184.
- Reiss, J. (2008) *Error in Economics: Towards a More Evidence-Based Methodology*. London: Routledge.
- Reiss, J. (2012). The Explanation Paradox. *Journal of Economic Methodology* 19(1). 43–62.
- Roubini, N. & Mihm, S. (2010). *Crisis Economics: A Crash Course in the Future of Finance*. London: Penguin.
- Sen, A. (1981). Accounts, Actions and Values: Objectivity of Social Science. In: Lloyd, C. et al. (eds.) *Social Theory and Political Practice*. Oxford: Oxford University Press
- Stiglitz, J. (2009). The Anatomy of a Murder: Who killed America's economy. *Critical Review* 21(2–3). 329–339.
- Stiglitz, J. (2010). *Freefall: America, Free Markets, and the Sinking of the World Economy*. New York: Norton.
- Stiglitz, J. (2011). Rethinking Macroeconomics: What Went Wrong and How to Fix It. *Global Policy* 2(2). 165–175.
- Strevens, M. (2007). Why Explanations Lie: Idealization in Explanation. Available from <http://www.strevens.org/research/expln/Idealization.pdf>.
- Sugden, R. (2009). Credible Worlds, Capacities and Mechanisms. *Erkenntnis* 70. 3–27.
- Teller, P. (2001). Twilight of the Perfect Model Model. *Erkenntnis* 55(3). 393–415.
- Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.
- Ylikoski, P. (2011). *Constitutive counterfactuals and explanation*. Manuscript. Tampere University.

### **Prof. Dr. Julian Reiss**

Durham University  
Department of Philosophy  
50 Old Elvet  
Durham DH1 3HN  
United Kingdom  
[julian.reiss@durham.ac.uk](mailto:julian.reiss@durham.ac.uk)

Peter König, Kai-Uwe Kühnberger, and Tim C. Kietzmann

# **A Unifying Approach to High- and Low-Level Cognition**

## **Introduction**

### **Cognitive science on low- and high-level – A divided land**

From its early beginnings to today, the interdisciplinary endeavor of cognitive science has led to a fundamentally improved understanding of many aspects of cognition. Some of this is due to the multi-leveled approach, as researchers have adopted a wide variety of techniques to understand cognitive phenomena at various levels of description. One way of distinguishing these different levels is by separating them into high- and low-level cognitive processes. Whereas the former includes cognitive abilities like planning and reasoning, the latter is generally seen as including the various modalities of sensory processing.

There are many reasons for such a seemingly principled division. For instance, low-level cognition, such as sensory processing, exists in virtually all animal-species, whereas high-level cognition, as described in more detail below, is mostly ascribed to human cognitive processing. In terms of bandwidth, vision, a low-level cognitive function, is the most dominant sensory modality in humans. Vision can be found in most species, specifically in all chordates (Land and Fernald 1992). In many of the latter, e.g. birds of prey, the spatial acuity of the visual system even surpasses human performance by a factor of 2 and more (Reymond 1985). Similar statements can be made with respect to other sensory modalities, such as audition. Sophisticated auditory systems are found in all chordates (Alexander 1981) and many species outperform human capabilities with respect to frequency range or sensitivity. Importantly, similarities across species can also be found with respect to the structures supporting sensory processing. For instance, although many different forms of receptors for optic signals can be found, lens-bearing eyes, as present in vertebrates, have evolved several times (Land and Fernald 1992; Nilsson 1989). Moreover, more proximal structures that support sensory processing exhibit similar organizational structures (Kaas 1997). From this it can be concluded that high performance sensory processing is a general capability, performed by most living species and that it is mostly based on related principles.

With regard to high-level cognition, there is no general definition or classification available and the typical assignment is mostly based on intuitions. Yet,

researchers agree that logical reasoning, planning and language belong to its core capabilities (Thagard 2008). On a broader scope, decision-making, memory skills, creativity, general intelligence and social interactions are also mentioned in this context. Contrary to low-level cognitive processes, these capabilities are mostly thought of as being uniquely human. As studies comparing human and animal performance are still scarce, reports of intelligent animal behavior are greeted with great attention (Watanabe ND Huber 2006; Blaisdell et al. 2006). Summing up, the present state of research ascribes high-level cognitive processes primarily to human cognition.

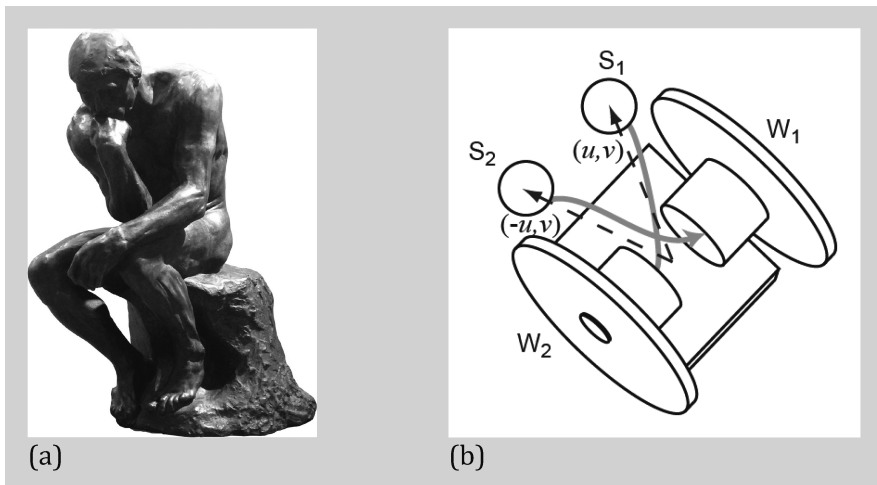
A second case for the view of a principled division can be made by investigating in how far low- and high-level cognitive processes are approachable with modern information processing techniques. Whereas visual processing in artificial systems, again classified as a low-level cognitive function, can greatly benefit from our increased understanding of the cortical visual system (Pinto et al. 2008; Kietzmann et al. 2009), high-level cognition poses more challenging problems. In the realm of sensory processing, ideas flow back and forth between the two disciplines and the performance of artificial systems can be quantified and compared to human performance. Importantly, common belief holds that there are no principled obstacles to achieving near-perfect performance with artificial systems. In contrast to this, state-of-the-art computer systems targeting high-level cognitive capabilities, as defined above, do not (yet) resemble neuronal structures in the least. Doctor Dostert predicted that “five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact”<sup>1</sup> (IBM 1954). After more than 50 years, with machines that are 100.000.000 times faster<sup>2</sup> problems originating from the domain of high-level cognition are still considered as extremely difficult, even though many of them are in fact considered simple by human standards. Impressive advances have been made in the context of well-defined artificial settings, e.g. chess playing, but artificial systems still perform poorly in high-level cognitive tasks that require a combination or integration of many specific high-level abilities. A good example is the usage of natural language, an ability that requires the integration of background knowledge, linguistic knowledge, reasoning, pragmatic aspects, gestures etc. Hence, the widely disparate progress approaching low- and high-level cognitive tasks in artificial systems underlines the view of a principled division.

---

<sup>1</sup> [http://www-03.ibm.com/ibm/history/exhibits/701/701\\_translator.html](http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html).

<sup>2</sup> <http://www.tomshardware.com/reviews/core-i7-990x-extreme-edition-gulftown,2874-6.html>, <http://www.ibm.com/ibm100/us/en/icons/ibm700series/>.

Finally, cognitive phenomena can be divided into high- and low-level on the basis of their relation to our body and actions in the real world. Sensory processing seems to be necessarily connected to actions performed by natural agents of all developmental levels in real environments (Varela et al. 1991). Moreover, it has been proposed that knowledge of changes of sensory signals contingent on performed actions is constitutive of perceptual consciousness (O'Regan and Noe 2001). In contrast to this, high-level cognition, at least in humans, allows for abstract reasoning processes that are completely decoupled from concurrent or subsequent actions. This line of reasoning does of course not imply that the development of abstract reasoning abilities can evolve without grounding in real world actions. Nevertheless, reasoning may be performed without direct reference to actions in the world, once they are developed. Rodin's "the thinker", being deep in thoughts while completely immobilized, can be seen as a symbolization of this decoupling of high-level cognitive processing and actions in the world (Figure 1a). Contrary to this, the Braitenberg vehicle (Braitenberg 1984, Bach et al. 2007) represents a direct interaction of sensory processing and corresponding actions (Figure 1b)<sup>3</sup> and is therefore an example of pure low-level cognition.



**Figure 1:** The difference between high- and low-level cognition can be symbolized by the contrast of "the thinker (Rodin)" and a Braitenberg vehicle.

<sup>3</sup> Figure 1b adapted from Bach (2009) with author's permission.

These arguments match the common intuitions of a fundamental gap between the stream of sensory input and the conceptual or symbolic level of interpretations. In fact, it still seems to be rather magic how analog and distributed sensory input can be “lifted” to a symbolic level on which many high-level cognitive processes operate and, back again, how reasoning processes can be propagated to the actuator level. Taken together, these properties argue for a principled division of labor between two different cognitive systems employing qualitatively different algorithms suited for their respective problem domain.

A question that arises naturally from this division of labor is whether the two systems, high- and low-level cognition, share common neural substrates or whether their function is in fact reflected in distinct cortical systems. Evidence for the latter is based on the long tradition of investigations of functional losses after localized cortical lesions. Here it was demonstrated that lesions to different parts of the cerebral cortex result in selective loss or changes in functionality of high- or low-level cognitive processing capabilities. For instance, damage to the cortical region in the occipital pole was found to lead to a loss of visual function and perceptual awareness. Additionally, to name just a few, blindsight results from a lesion of early visual areas (Weiskrantz 1968, Stoerig and Cowey 1997); visual agnosia was shown to be the result of lesions to the occipitotemporal cortex (Farah 2000); prosopagnosia can be elicited by lesions to the inferior occipital cortex or fusiform gyrus (Steeves et al. 2009); akinetopsia by lesioning parietal cortex (Zihl et al. 1983); achromatopsia by lesions to the ventral occipital cortex (Zeki 1990); and personality and behavior were shown to be affected by lesions to the frontal lobe (Barker 1995). Today, this view is complemented by studies applying modern imaging techniques, such as fMRI, PET, EEG, and MEG that demonstrate a functional compartmentalization in far reaching areas of cognition. Indeed, a fair part of the research effort focuses on the localization of cognitive functions and the remaining white spots on the cortical map are shrinking quickly. Thus, in addition to above observations, also the rapidly growing number of experiments that identify various functional specializations of different cortical regions speak in favor of a cortical dissociation between high- and low-level cognition.

## **Re-unification by a statistical approach and embodiment**

From the apparent differences in high- and low-level cognitive processes it could be proposed that both need to also rest on qualitatively different cortical opera-

tions and that the respective functional modules therefore exhibit structural differences. Indeed, individual areas have been delineated based on cortical structure in the form of lamination (Brodman 1909). Complementing this structural approach, more recent imaging techniques have been used to provide *functional* definitions of cortical modules (Felleman and van Essen 1991; Hilgetag and Barbas 2009). By taking both types of information together, it is now possible to validate the claim of a direct structure-to-function mapping. As a result, mostly early sensory areas and the primary motor areas were shown to be dissociable based on functional as well as structural information. However, many regions that can be functionally separable cannot be distinguished based on their anatomical structure. Among others, this holds for the large variety of areas in the intraparietal lobulus (LIP, VIP, MIP, PRR, AIP), which were shown to be functionally distinguishable despite all being situated within Brodman area 39. In addition to this structural similarity across different functionally defined areas, larger scale structures exhibit functional restructuring. This highly impressive capability of the human cortex was demonstrated for auditory information that can successfully be rerouted to visual cortex (Sur et al. 1988). It is also evident in blind subjects for whom the visual cortex seems to fulfill detailed sensory information processing during Braille reading (Sadato et al. 1996; Hamilton and Pascual-Leone 1998; Merabet et al. 2009). From this we can conclude that cortical modules are not limited to their primary function but can adapt to a wide variety of tasks. Importantly, it can be hypothesized that the quantitative properties typically used for a structural separation are more related to a fine-tuning of function, but not to qualitative differences of operations in the respective areas. That is, different functional specializations do not necessarily match one-to-one on different structural specializations. This allows for the proposition that operations performed in different cortical modules, including both high- and low-level cognition, are also not as distinct as the supported functions might suggest, but that they are in fact rather comparable. This resonates well with the concept of a canonical microcircuit (Douglas and Martin 2004), which holds that the structure of neocortical circuits is general and that neuronal circuits in neocortex can therefore be considered canonical.

Given these observations, we hypothesize that a similar approach can be taken for the description of high- and low-level cognition: Although both are based on different networks of functionally defined cortical regions, both types of cognitive processing may in fact implement comparable operations. As a result, *functional differences arise solely from different statistical properties of afferent signals and different context of those networks and not from inherently different structural properties*. Put differently, we argue that low-level sensory processing has many more similarities with high-level cognitive reasoning than previously

assumed. To illustrate this admittedly bold claim, we concentrate on two central examples of low- and high-level cognition in the remainder of this article: invariant object recognition and analogical reasoning.

## **An example of a low-level cognitive process: object recognition**

### **The hard problems of object recognition**

Cognitive tasks of diverse complexity rely on a successful and reliable recognition of objects. For instance, consider the recognition of your car in a parking lot. Without problems you can identify it immediately in a large array of similar objects, despite different light-conditions, occlusions, different viewpoints depending on the direction in which you approach it, different retinal sizes that arise from different distances, as well as largely diverse background colors and clutter. Importantly, object identification is fast. Thus, even if timing is more crucial and the environment is more dynamic, as in the case of being part of a soccer game, we are immediately able to recognize the ball, independently of its color or texture, together with the goal and other players albeit their dramatic changes in shape upon movement. Finally, consider the more general case of object classification, as for instance in the case of classifying an animal as a dog. Despite the large variety of sizes, colors and types of dogs, we are very well able to successfully complete this task.

As these examples illustrate, both types of object recognition (identification and classification) belong to the most essential capabilities of the human visual system and prepare the grounds for many higher-level cognitive processes. Although we perform this task constantly and seemingly without effort, it is an extremely difficult problem from a computational point of view, as exemplified in above examples.

## **Divide and conquer: processing in the visual hierarchy**

How is the visual system set up in order to solve this complicated task? What cortical structures enable the system to fulfill the requirement of highly specific



and at the same time robust classifications, i.e. to solve the specificity-vs-invariance problem? In the human brain, invariant object recognition is largely accomplished in the ventral visual stream (Haxby et al. 1991), which exhibits a hierarchical structure (Felleman and van Essen, 1991). Starting from retinal input, which passes through the Lateral Geniculate Nucleus (LGN), information enters striate cortex (V1), in which neurons are selective to bars of light and basic colors. Further downstream, information passes through areas in which neurons exhibit receptive fields of increasing complexity and size. These include the areas V2 and V4, in which color constancy is accomplished, and the lateral occipital complex, which is selective to spatially congruent, informative object parts (Lerner et al. 2008). Finally, information enters the cortical structures in the inferotemporal cortex. Here, cells exhibit selectivity for complex shapes including selective object views and view-invariant object representations (Tanaka 1996). Moreover, recent work demonstrates that neurons in the medial temporal lobe combine high selectivity for individuals with impressive invariance operations (Quiroga et al. 2005) and it is indeed possible to reliably classify and identify visual objects based on small populations of neurons in inferotemporal cortex (Hung et al. 2005). These results, and many more that cannot be covered in the scope of this article, paint the picture of a systematic division of labor within the ventral stream of the visual system. Following the hierarchy of visual areas, neurons exhibit more and more complex and at the same time increasingly robust response properties that lead to representations suitable for explicit object recognition.

Despite our increased understanding of the different neuronal selectivity across the ventral stream, however, it remains largely an unsolved question which principles underlie the development of these hierarchical representations and underlying cortical structures. Here, the class of normative approach is particularly promising, as detailed below.

## **Optimality as a general statistical principle: from sparseness to stability**

In recent years, an increasing number of studies explicitly addressed the variability of neuronal response properties by a normative approach. This notion dates back to Barlow's fundamental principle that neuronal representations should comply with the relevance for the animal, be suitable for decoding by downstream areas, and allow for efficient encoding by virtue of redundancy reduction (Barlow 1961). Specifically the latter endorses the approach that sensory processing should optimize mathematically defined criteria. These criteria are optimized

for a given set of inputs, i.e. natural sensory stimuli. This rather different approach towards understanding sensory processing has mostly been studied in the visual domain, where computational models have successfully demonstrated the emergence of receptive field types exhibiting neuronal properties that are comparable to the ones found in the visual cortex. Notably, the normative approach is complementary to the experimental approaches: rather than measuring response properties of neurons in individual cortical areas, they are understood as the effect of unsupervised learning from natural input and its statistics.

The normative approach, which presupposes that neuronal representations optimize an objective function, requires a definition of the target properties. Following the requirement of efficient coding, optimality was first formalized on the basis of sparse representations. This implies that each representational unit specifically codes only for a small subset of the typical stimuli; i.e. neuronal receptive fields should be shaped in a way such that they lead to action potentials for only a small set of effective stimuli. Indeed, it has been found that the neuronal selectivity in the area V1 can be understood as adhering to a sparse code, given natural input (Olshausen and Field 1996). It was shown experimentally that natural stimuli evoke sparse activity patterns not only in V1, but also in higher visual cortices. Moreover, the application of a sparse coding scheme to intraareal interactions leads to functional coupling that is compatible with the lateral connectivity in V1 (Garrigues and Olshausen 2008). Finally, optimally sparse representations are closely related to independent component analysis, a statistical method suitable to infer the independent sources of a superposition of signals (Bell and Sejnowski 1997; Hyvärinen and Oja 2000).

Sparse coding leads by definition to high levels of specificity. This is due to the fact that sparseness enforces representations that react to only a small fraction of possible input. As seen in our earlier examples, however, specificity does not suffice for successful and robust object recognition, as invariance to sensory fluctuations and viewing conditions is equally important. This idea is picked up in another family of coding principles, which is based on the temporal continuity of natural stimuli. Despite changing implementations and names (stability, slowness, temporal coherence, etc.), the underlying assumption of these approaches is that relevant properties typically vary on a slower time-scale than irrelevant ones (Földiak 1991; Körding and König 2001; Wiskott and Sejnowski 2002). Thinking back to our first object recognition examples above (your car in a parking lot), what is common to all of the described complications is the fact that although sensory sampling differs largely from one situation to another, the identity of the object remains constant. Exactly this observation is capitalized upon with temporal coding schemes, which imply that the identity of an object changes on a slower timescale than the associated sensory information. Again, correspond-

ing computational simulations targeting at striate cortex proved to be rather successful (Wiskott and Sejnowski 2002, Einhäuser et al. 2002; Körding et al. 2004, Berkes and Wiskott 2005). It was shown that stability does not only lead to simple-cell-like receptive field structures, but also that it can explain the phase invariance of complex cells. Although many questions are still open (Olshausen and Field 2005) the normative approach has led to significant progress in a principled understanding of primary visual cortex.

As an obvious next step, the normative approach was extended to higher visual cortices further down the ventral stream. For instance, it was shown that a stability-optimizing neural network increased the rotation invariance of the emerging representations, thereby enhancing recognition capabilities in a set of readout neurons (Einhäuser et al. 2005). Moreover, simulations of hierarchical networks based on the visual input of an artificial agent in a natural environment demonstrated the emergence of increasingly complex, yet stable visual representations. At the upmost hierarchical level, higher-level representations were shown to emerge that were selective to the position of the agent in space, but invariant with respect to its orientation (Wyss et al. 2006; Franzius et al. 2007). These matched properties of place cells as observed in the hippocampus (O'Keefe and Dostrovsky 1971). Finally, cells responsive for head-direction and spatial view-cells can be explained by the same set of principles (Franzius et al. 2007, Sprekeler and Wiskott 2011).

In addition to the computational work, important support for the stability approach was provided by electrophysiological experiments in which it was demonstrated that targeted changes of the temporal contiguity of objects lead to changes in response properties in inferotemporal neurons – a direct prediction of a neuronal coding scheme that is based on the stability principle (Li and DiCarlo 2010). These important results demonstrate that the normative approach does not only give a faithful description of neuronal response properties throughout the ventral stream, but that it also predicts the consequences of experimental manipulations. Moreover, the hierarchical application of the stability principle is a promising candidate in the attempt to close the gap to invariant object recognition. Thus, invariant object recognition and the development of neuronal response properties can be partly understood as a consequence of optimal sensory representations.

Besides to the well-studied visual domain, the normative approach has also been applied to other sensory modalities, such as auditory and somatosensory processing (Klein et al. 2003; Hipp et al. 2005; Duff et al. 2007). Taken together, these studies indicate that the wide variety of response properties on different levels of the visual system and of other modalities are fully compatible with a

single set of principles governing sensory processing: Sparseness, slowness and decorrelation.

Above considerations are mostly based on the case of object recognition in which variations in sensory sampling originate from one object. However, it can also be argued that the resulting networks exhibit a most crucial new property: they can generalize from invariant object identification to the case in which different objects are associated with one label (object classification, our third example above). If object classification is understood as requiring invariance over object identities (as opposed to sensory variation), then the task could in principle be accomplished by the same normative approach as that lead to an increasing invariance over sensory sampling only. If this is indeed possible, then this implies that different objects of the same category share similar aspects of cortical representation. It has to be noted, however, that despite our ability to classify objects, we are still well able to recognize individual object instances under a great variety of conditions and viewpoints.

## **Combining supervised and unsupervised learning schemes for successful object recognition**

The renaissance of neural networks in the '80s of the 20<sup>th</sup> century is tightly linked with the discovery and re-discovery of training methods for hierarchical neural networks (Werbos 1974; LeCun 1986; Rumelhart et al. 1986). How does an unsupervised training scheme, such as the normative approach described above, match with the typically utilized supervised algorithms of artificial neural networks? For the latter, the parameters and connection weights are iteratively tuned to match the output to the desired result. By now this work has expanded to a huge field and excellent reviews and textbooks are available (Bishop 2006). For the present purpose, however, we want to highlight a single specific problem only. These supervised learning procedures require labeled data, which are scarce and expensive in real life and thus might hinder proper convergence and generalization of the network structures. Hence, it is attractive to combine these methods with unsupervised learning, i.e. a normative approach described above. Indeed, applying unsupervised training to all layers of a hierarchical network but the last, and complementing this approach with supervised training of the output layer significantly reduces complexity of learning at a small price in performance only (Einhäuser et al. 2005, Franzius et al. 2008). Hence, the normative approach is fully compatible and fosters object recognition in hierarchical networks.

## From optimal sparseness, and optimal slowness to optimal predictability

The family of temporal coherence/slowness/stability approaches has been shown to explain many aspects of receptive field properties found in the visual hierarchy and thereby provides a principled approach for understanding invariant object recognition. However, if stability based on the statistics of natural input was the only objective function that is optimized in the mammalian cortex then the question arises why different species exhibit radically different sensory representations. Previously, we have put forward the hypothesis that sensory systems optimize the capabilities to predict and support sensory consequences of actions (König and Krüger 2006). Moreover, sensory selectivity should be shaped in a way such that they optimally support the potential actions of the agent. Because of this, neuronal representations should also be tuned to address those features that are optimally predictable with respect to the agent's actions. This entails the crucial step that the sensory predictability is integrated into the previously defined objective function (Weiler et al. 2010).

With the reference to different actions, the principle of predictability refers implicitly to the behavioral repertoire of the agent. Compared to the previously mentioned principles, this is a decisive step. Given that the visual systems of humans, non-human primates and carnivores differ in profound ways, relating visual processing to the behavioral repertoire opens a new avenue to understanding these differences.

## Towards optimal high-level processes: the example of analogies

How can we bring together the normative approach, which has been successfully applied in the domain of visual processing, a low-level cognitive function, with a high-level cognitive process such as the formation of analogies? In this section we will present our central claim that the principle of optimal action predictability and invariant actions supplies a unified framework of low-level and high-level cognitive functions. With this concept, we move from investigating purely sensory features to active representations that jointly address sensory information and the agent's action repertoire.

As an illustrative example, consider the case of the soccer ball from above. Kicking the ball requires the player to first recognize the individual ball, an item

which was trained earlier, before any aiming or kicking can be accomplished (object recognition). However, if we were to swap the ball with a different one, it would nevertheless be possible for the player to immediately recognize the item to be kicked and to perform the appropriate action. Although this example might seem trivial at first, the performed computation is more complicated. This is because the player did not only generalize from one ball to a different exemplar (classification based on afforded actions), but also performed a generalized prediction of the consequences of the action. It is therefore an example of invariant actions. Importantly, despite the simplicity of the example, what has happened through the described generalization in sensory-motor space can in fact be seen as the drawing of an analogy. In the following sections, we will first describe the general research on (predictive) analogies before we describe the details on how both research areas can be understood on the basis of a unifying approach.

## (Predictive) analogies

Analogies are in the intense focus of research addressing high-level cognitive processes. Although an important topic in many disciplines for a long time, their scientific study in the context of cognitive science started with the seminal paper (Gentner 1983) introducing the *Structure-mapping theory*. This theory is based on the idea that establishing an analogical relation is a structural comparison of two domains, such that an “interesting” substructure in the source domain is aligned to an “interesting” substructure of the target domain. In other words, the forming of analogies relies on identifying commonalities of the two substructures. The structure-mapping theory has been proven its remarkable potential and is a de facto standard in cognitive models of analogy-making. Furthermore, the technical realization of the structure-mapping engine provides a standard computational model (Falkenhainer et al. 1998).

There are at least three classical domains from which typical examples of high-level cognition involving analogical reasoning are drawn: geometry (Evans 1968), naïve physics (Falkenhainer et al. 1998), and formal languages (Hofstadter et al. 1995). Besides these classical domains, however, many other domains have been discussed in the literature (intelligence tests, metaphorical expressions of natural language, problem solving, didactics of mathematics, sketch recognition etc.). Accompanying the variety of domains, researchers proposed a variety of different frameworks to account for the observed phenomena, ranging from symbolic models, like *Structure-mapping theory* or *Heuristic-driven-theory-projection* (Schwering et al. 2009), to neurally inspired frameworks, like *Learning and*

*inference with schemas and analogies* (Hummel and Holyoak 1997) and hybrid approaches, like *Associative memory-based reasoning* (Kokinov and Petrov 2001). Despite its symbolic nature, *Heuristic-driven-theory-projection* and *Structure mapping theory* explicitly distinguish between low-level and high-level processes. *Associative memory-based reasoning* models all cognitive levels, but explicitly distinguishes between symbolic (reasoning-related) representations and neurally inspired activation spreading for attention and priming mechanisms. A similar separation holds in our opinion for the concept of *Learning and inference with schemas* and analogies. Hence, all of these frameworks accept the principled division of low-level and high-level cognitive processes.

An important class of analogies is given by so-called predictive analogies (Indurkha 1992). Predictive analogies explain a new domain (target) by transferring information (knowledge) from the source to the target, such that non-trivial new conclusions can be drawn in the target domain. Because of this productive aspect, (predictive) analogies are often considered as a source of creativity and a mechanism for analogical, i.e. concept-guided, learning (Friedmann et al. 2009; Gust and Kühnberger 2006). For example, in the naïve physics domain, predictive analogies relate physical domains that are hardly accessible by our direct experience to domains that have perceivable properties. Due to the rich explanatory power supported by the source domain it is therefore possible to draw predictions in the target domain, which can then be experimentally evaluated. For example, let's consider an analogy between a water pipe system and an electric circuit. In the water pipe system, it can be observed that a “current” is triggered by “pressure” and that the system is necessarily closed. Another observation would be that “narrowing wires” influence the ongoing current. If we now apply these observations to previously learned concepts from the domain of electricity (the flow of electrons is triggered by a voltage difference, and a resistor influences the flow of electrons), the analogy is striking. Importantly, the analogy allows for the possibility of drawing new inferences. An example of such a prediction would be that adding a further resistance into the circuit should again reduce the flow of current. Notice that although these two exemplary domains were both chosen from the field of naïve physics, they do not show strong similarities, but are quite different from each other concerning the observable properties. Yet, the formation of analogies allowed for predictions from one domain to another.

## Object classification as predictive analogy

Now let us consider predictive analogies in the context of object recognition. First, it should be noted that analogies are already used in current visual sketch recognition systems and in systems designed for the recognition of geometric regularities in intelligence tests (Lovett et al. 2009; McLure 2010). Although such applications of analogy-making systems in the field of object recognition are rather new, some promising results for standardized tests in the geometry domain have already been achieved (e.g. Raven's progressive matrices; Lovett et al. 2010). Sketch-recognition systems for analogy-making typically work solely on the perceptual level, i.e. they identify important regions and features of such regions for an analogical comparison. In contrast to this, further aspects of cognition, like possible actions or functional properties of objects that are represented, do not play a role.

As an additional step, let us consider a case in which possible actions or action-outcomes can be included. Let us again consider the case in which we are presented with an object (e.g. a soccer ball) and need to classify it. Starting with the visual features of the object, it is possible to deduce properties that are relevant for an interaction with the object. We call this a 'predictive property' of an object. An example of such a property is "when kicked, it will roll". Furthermore, we might reasonably expect "when rolling on flat smooth ground, it will continue to roll for some time". However, this sequence of processing steps (from visual to predictive properties) is not necessarily required. Instead of starting with a visual analysis leading up to the recognition of an object and onwards to potential actions, we can also twist the argument and assign the primary importance to potential interactions with the object and thereby base object recognition on this set of afforded actions. Now, an object that satisfies the properties "when kicked, it will roll" and "when rolling on flat smooth ground, it will continue to roll for some time" is by (functional) definition a ball. With this, we have moved from a purely visual to a functional definition. Nevertheless, the provided functional definition of a ball can still be fulfilled via purely visual properties.

With the above case of a soccer ball, we have intentionally chosen an introductory example that is highly suitable for the classical approach that starts with a visual analysis that leads up to object recognition and only from there to functional predictions and we presented the functional definition as an alternative view. A visual definition can be seen in the well-known tradition in linguistics and logic of defining a concept by its intention, i.e. by the properties and attributes of the concept (Frege 1960). In the visual domain, such properties and attributes must be perceivable and as discriminative as possible. Important differences between this tradition and our proposal to define a ball functionally are



the explicit emphasis of action-centered and manipulatory properties and the predictive character of such properties in the functional definition. Yet, already a slightly more complex example uncovers numerous problems with this purely visual approach. What is a chair? A quick look into Wikipedia gives a reasonable description: “a chair is a stable, raised surface used to sit on, commonly for use by one person”<sup>4</sup>. In normal circumstances, a raised surface can be defined based on visual features. A chair put upside down, however, has no raised surface anymore and thereby violates the visual definition – yet it is still a chair. On the other hand, a cube does have a raised surface, but it is usually not considered to be a chair. These problems leave us with the remainder of the definition: “to sit on”. This part puts the focus on the use of the object and is in essence a predictive analogy of the form “when you put your weight on it, you will not fall down”. In line with this, the Oxford Dictionary directly concentrates on the function of a chair, which is “a separate seat for one person, ...”<sup>5</sup>. Notice that such problems occur necessarily with every intentional definition, because it is not possible to give a sufficient and necessary set of (visual) properties and attributes for classifying every potential instance correctly. Thus, as an alternative to a purely visual definition, we follow the approach that predictive analogies map functional connections and thereby are a vital part of the object definition.

While it has to be admitted that the original definition of *predictive* analogies does not perfectly fit into the domain of object recognition, successful performance in the visual domain nevertheless requires a transfer of (functional) knowledge from known examples to unseen ones in order to make the right classification and to select appropriate actions. Hence, although the original context of predictive analogies is in fact a different one, it seems unproblematic to call such analogies in the domain of object recognition to be predictive.

## How can we understand the emergence of analogies? – A unifying approach

Again, we start with a simple example: the case of driving your car to work in the morning and back again in the evening. Of course, while approaching your car in the morning, you recognize it albeit visually very dissimilar conditions. Moreover, (higher) cortical areas implement afforded actions (cyan). Once in the car, a specific action representation is activated (blue region) when you press the

---

<sup>4</sup> <http://en.wikipedia.org/wiki/Chair>.

<sup>5</sup> <http://www.oxforddictionaries.com/definition/chair?view=uk>.

right lever (gas pedal) with your foot and the sensory consequences of this action are predicted. This situation is visualized in Figure 2 in which part of the sensory representation (green area), as well as the action state (blue area, e.g. extension of the right leg and foot represented by changes in nodes 10 and 14) are altered by the active afforded actions (node 12 in cyan area). Because the now altered state, the afforded actions have changed as well (cyan area, pushing the break would now have an effect and opening the door is prohibited as represented by changes in activity of nodes 9 and 13). After work on the way home, it can be assumed that large parts, although not all, of active visual representations are identical to the ones activated in the morning<sup>6</sup>, while some aspects might differ (e.g. no coffee in the cup holder (green node 7)). This has consequences on the sensory, afforded actions, and motor level. Yet, pressing the right pedal yields as expected the same effect and the sensory representation is transformed. In this case, the analogy is supported by largely overlapping sensory and action representations, which is in turn due to the performed invariant object recognition.

Compared to your own car, driving a different car to work introduces some more changes. For instance, the color and the geometry of the seat might be different. Yet, pressing the right pedal does lead to an acceleration and the predictions of sensory changes based on the experience on the former car are correct. This again can be considered as an analogy, which is supported by invariant object classification (different individual, same class) and its associated predictive properties. Indeed, we argue that despite an overwhelming amount of variance of sensory signals, the basic functionality is identical in both cases. Only because of this can the two objects be considered to belong to the same class. In this case the problem of invariant action representation has been transferred to invariant object classification. Now again consider driving a pellet jack. This might yield unexpected results although it does have a steering wheel, foot pedal and is part of the general category car. Yet, some come with a left foot accelerator pedal and using it in the usual way leads to a mismatch of predictions of the sensory consequences of actions and reality. Although the object is obviously a vehicle allowing a partly overlapping set of afforded actions (including pressing the right pedal), the result is not the same and the analogy breaks down. This demonstrates that the predictive framework is in fact working on probabilities and thus does not always allow for literal logical inferences.

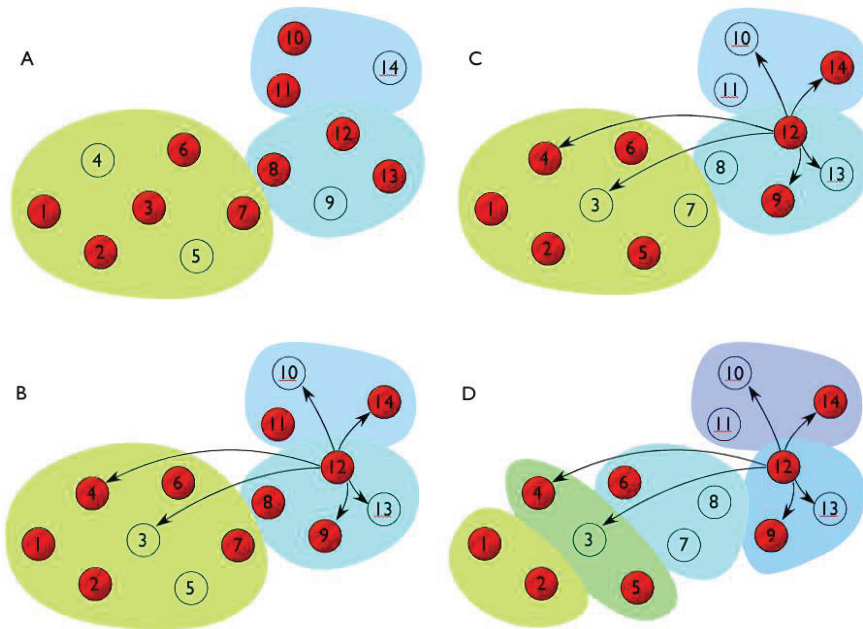
Although our examples might lead to this view, it is in general not possible to neatly divide the different representational areas into sensory representa-

---

<sup>6</sup> For the sake of the argument and visualization we assume a highly sparse representation of sensory signals (green). Please note, however, that a coarse population code does not change the principle of the argument.

tions, afforded actions, and motor representations (figure 3D). To the contrary, a gradual transformation of sensory representations to motor actions leads to a setup in which neurons at each level act as “sensory” representations feeding bottom-up input to higher levels, modulate other potential affordances by tangential interactions at the same level and predict “action” induced changes of sensory representations at the lower level. The label of sensory/affordance/motor representations is therefore relative to the viewpoint. Still, even with this view the approach is fully compatible with a hierarchical network composed of general processing units according to a small set of optimization principles.

In our car example, the common set of afforded actions directly mirrors the invariant processing in the bottom-up pathway and sensory representations can be assumed to largely overlap. Hence, the prediction of sensory changes induced by the afforded action applies to the whole set of similar sensory representations. This is, however, not a necessary precondition. In general the afforded action is dependent only on a small part of the sensory representation and invariant with respect to other parts. This property defines it as an invariant action, which is at the core of making a predictive analogy.



**Figure 2:** Schema of gradual transformation of sensory representation via affordances to motor actions. For detailed description see text.

## Easy and difficult analogies

In many everyday situations (and in the example above), the analogy comes quite natural such that, although being one, it is often not considered to be an analogy at all. In this section, we now consider more complex situations in which the overlap of sensory representations is not that large and the concept of invariant actions is more explicit. In classic examples, the basic constituents differ from each other in fundamental ways: To see this, consider, for example, the famous Rutherford analogy between the solar system, i.e. a system of planets revolving around a sun, and an analogous atom model, in which electrons are no longer homogeneously distributed as in the historically prior “plum pudding” model, but are revolving around a nucleus. In such examples, the overlap of sensory representations is minimal or even not existing and structural commonalities on a higher, i.e. abstract, level seem to be important. This brings us back to the origins of the scientific study of analogies in cognitive science in which rather abstract domains were considered. How are analogies emerging under such conditions that seem to be completely decoupled from any sensory representations? An explanation can be given by considering a situation in which the solar system – atom model analogy is visualized in form of diagrammatic representations (as in a scenario of a teaching situation in high school). In this form of representation, the constituents are in fact very similar to each other. There is a center, revolving objects represented as little circles and there are attracting and retracting forces etc., in short, the analogy is striking. It is rather uncontroversial that the emergence of the abstract conceptualization of a revolution movement is without any doubts grounded on a simpler, more concrete level and learned by using simpler, more concrete examples. Sensory representation, among other aspects as, is such a concrete level establishing a solid foundation of such generalizations.

## Object recognition, context, and actions

It is well-known that object recognition performances in psychological experiments change significantly, if the object in question is put into varied more or less prototypical contexts (flying eagle vs. sitting eagle, Zwaan et al. 2002). If object recognition has anything to do with establishing analogical relations, then context effects need to be considered also for analogies. How can context effects be transferred to the domain of analogy making? We suggest that contextual effects in the visual domain are quite often reducible to even more fundamental afforded actions and their representations. This is mainly due to the fact that per-

ception and recognition tasks never occur without a dynamic environment and an active agent. For natural scenes, the coupling of the recognition of an object and action-related aspects is natural (eagles fly, planes fly as well, therefore they need wings etc.).

## Summary and conclusions

We propose that a key to overcome the artificial separation of low-level and high-level cognition is the concept of invariant actions, which optimally predicts action-induced changes of sensory signals. This concept is rooted in the ideas of Gibson (1977), yet makes crucial extensions. (1) To a first order of approximation, cortical processing is based on cortical modules of homogeneous structure. The function of these modules is to transmit optimally predictable parts of afferent signals to higher levels and to make predictions of changes of lower-level representations. Hence, functional differences originate mostly in the differences in input/output connectivity. (2) The optimization process leads to the emergence of invariant representations. Afforded actions emerge gradually in a hierarchical processing scheme obviating a strict separation in sensory and motor representations. Focusing on the bottom-up direction, these might be viewed as invariant object representations, focusing on the top-down direction, these are invariant action representations. (3) Invariant actions are the core of predictive analogies. In most situations, the invariance is so natural that we emphasize invariant object recognition and do not realize that the implied actions are based on predictive analogies. The more arcane invariant actions, the classical examples of predictive analogies, are at the heart of higher cognitive functions. Together, these three steps establish “optimally predictive active representations” as a unified description and postulate a uniform cortical substrate and functional mechanisms for low-level and high-level cognitive processes.

## References

- Bach, J., Bauer, C., & Vuine, R. (2007). MicroPsi: Contributions to a Broad Architecture of Cognition. *KI 2006: Advances in Artificial Intelligence. Lecture Notes in Computer Science* 4314. 7–18.
- Barker, F. G. (1995). Phineas among the phrenologists: the American crowbar case and nineteenth-century theories of cerebral localization. *Journal of Neurosurgery* 82. 672–682.
- Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In: Rosenblith, W. A. (ed.). *Sensory Communication*. Cambridge: MIT Press. 217–234.

- Bell, A. J. & Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision Research* 37. 3327–3338.
- Berkes, P. & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision* 5(6). 579–602.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
- Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science* 311(5763). 1020–1022.
- Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.
- Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig: Johann Ambrosius Barth Verlag.
- Douglas, R. J. & Martin, K. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience* 27. 419–51.
- Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics* 93(1). 79–90.
- Einhäuser, W., Kayser, C., König, P., & Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience* 15(3). 475–86.
- Evans, T. (1968). A program for the solution of a class of geometric-analogy intelligence-questions. In: Minsky, M. (ed.). *Semantic Information Processing*. Cambridge, MA: MIT press. 271–353.
- Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 20(41). 1–63.
- Farah, M. J. (2000). *The Cognitive Neuroscience of Vision. Fundamentals of Cognitive Neuroscience*. Malden, MA: Blackwell Publishers.
- Felleman, D. J. & Essen, D. C. van (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1(1). 1–47.
- Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computing* 3. 194–200.
- Franzius, M., Sprekeler, H., & Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLOS Computational Biology* 3(8). e166.
- Franzius, M., Wilbert, N., & Wiskott, L. (2008). Invariant Object Recognition with Slow Feature Analysis. In: *Artificial Neural Networks – ICANN 2008. Lecture Notes in Computer Science* 5163. 961–970.
- Frege, G. (1960). On Sense and Reference. In: Geach, P. & Black, M. (eds.). *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Basil Blackwell.
- Friedman, S., Taylor, J., & Forbus, K. (2009). Learning Naïve Physics Models by Analogical Generalization. In: Kokinov, B., Holyoak, K., & Gentner, D. (eds.), *New Frontiers in Analogy Research. Proceedings of the Second International Conference on Analogy*. Sofia: New Bulgarian University Press. 145–154.
- Freiwald, W. A. & Tsao, D. Y. (2009). Cingulate cortex: diverging data from humans and monkeys. *Trends in Neurosciences* 32(11). 566–574.
- Garrigues, P. & Olshausen, B. A. (2008). Learning Horizontal Connections in a Sparse Coding Model of Natural Images. In: Platt, J. C., Koller, D., Singer, Y., Roweis, S. (eds.), *Advances in Neural Information Processing Systems* 20. Cambridge, MA: MIT Press. 505–512.

- Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science* 7(2), 155–170.
- Gibson, J. J. (1977). The Theory of Affordances. In: Shaw, R. & Bransford, J. (eds.), *Perceiving, Acting, and Knowing*. New York: Wiley & Sons.
- Gust, H. & Kühnberger, K.-U. (2006). Explaining Effective Learning by Analogical Reasoning. In: Sun, R. & Miyake, N. (eds.). *28<sup>th</sup> Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 1417–1422.
- Hamilton, R. H. & Pascual-Leone, A. (1998). Cortical plasticity associated with Braille learning. *Trends in Cognitive Sciences* 2(5). 168–174.
- Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Schapiro, M. B., & Rapoport, S. I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences USA* 88(5). 1621–1625.
- Hilgetag, C. C. & Barbas, H. (2009). Sculpting the Brain. *Scientific American* 300. 66–71.
- Hipp, J., Einhäuser, W., Conradt, J., & König, P. (2005). Learning of somatosensory representations for texture discrimination using a temporal coherence principle. *Network: Computation in Neural Systems* 16. 223–238.
- Hofstadter, D. & the fluid analogies research group (1995). *Fluid Concepts and Creative Analogies*. New York: Basic Books.
- Hummel, J. & Holyoak, K. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104. 427–466.
- Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science* 310(5749). 863–866.
- Hyvärinen, A. & Oja, E. (2000). Independent Component Analysis: Algorithms and Application. *Neural Networks* 13(4–5). 411–430.
- Indurkha, B. (1992). *Metaphor and cognition*. Dordrecht: Kluwer.
- Kaas, J.H. (1997). Topographic maps are fundamental to sensory processing. *Brain Research Bulletin* 44(2), 107–112.
- Kietzmann, T. C., Lange, S., & Riedmiller, M. (2009). Computational Object Recognition: A Biologically Motivated Approach. *Biological Cybernetics* 100. 59–79.
- Klein, D. J., König, P., & Körding, K. P. (2003). Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing* 3. 659–667.
- Kokinov, B. & Petrov, A. (2001). Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In: Gentner, D., Holyoak, K., & Kokinov, B. (eds.). *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.
- König, P. & Krüger, N. (2006). Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics* 94(4). 325–334.
- Körding, K. P., Kayser, C., Einhäuser, W., & König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli. *Journal of Neurophysiology* 91(1). 206–212.
- Körding, K. P. & König, P. (2001). Neurons with two sites of synaptic integration learn invariant representations. *Neural Computing* 13(12). 2823–2849.
- Land, M. F. & Fernald, R. D. (1992). The Evolution of Eyes. *Annual Review of Neuroscience* 15. 1–29.
- LeCun, Y. (1986). Learning Processes in an Asymmetric Threshold Network. In: Bienenstock, E., Fogelman-Soulié, F., & Weisbuch, G. (eds.). *Disordered systems and biological organization*. Les Houches, France: Springer. 233–240.



- Lerner, Y., Epshtein, B., Ullman, S., & Malach, R. (2008). Class information predicts activation by object fragments in human object areas. *Journal of Cognitive Neuroscience* 20(7). 1189–1206.
- Li, N. & DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 67(6). 1062–1075.
- Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. *Proceedings of Cognitive Science* 10.
- Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive Science* 33(7). 1192–1231.
- McLure, M., Friedman, S., & Forbus, K. (2010). Learning concepts from sketches via analogical generalization and near-misses. In: Ohlsson, S. (ed.). *Proceedings of the 32nd Annual Conference of the Cognitive Science Society (CogSci)*. Portland, OR: Curran Associates, Inc.
- McNeill Alexander, R. (1981). *The chordates*. 2nd edition. Cambridge: Cambridge University Press.
- Merabet, L., Battelli, L., Obretenova, S., Maguire, S., Meijer, P., & Pascual-Leone, A. (2009). Functional Recruitment of Visual Cortex for sound encoded object identification in the Blind: A TMS Case Study. *NeuroReport* 20(2). 132–138.
- Nilsson, D.-E. (1989). Vision Optics and Evolution. *BioScience* 39. 298–307.
- Olshausen, B. A. & Field D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381. 607–609.
- Olshausen, B. A. & Field, D. J. (2005) How close are we to understanding v1. *Neural computation* 17(8). 1665–1699.
- O'Keefe, J. & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research* 34(1). 171–175.
- O'Regan, J. K. & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24(5). 939–973.
- Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLOS Computational Biology* 4(1). e27.
- Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435(7045). 1102–1107.
- Reymond, L. (1985). Spatial visual acuity of the eagle *Aquila audax*: a behavioural, optical and anatomical investigation. *Vision Research* 25. 1477–1491.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323. 533–536.
- Sadato, N., Pascual-Leone, A., Grafman, J., Ibañez, V., Deiber, M. P., Dold, G., & Hallett, M. (1996). Activation of the primary visual cortex by Braille reading in blind subjects. *Nature* 380(6574). 526–528.
- Sprekeler, H. & Wiskott, L. (2011). A theory of slow feature analysis for transformation-based input signals with an application to complex cells. *Neural Computation* 23(2). 303–335.
- Steeves, J., Dricot, L., Goltz, H., Sorger, B., Peters, J., Milner, D., Goodale, M.-A., Goebel, R., & Rossion, B. (2009). Abnormal face identity coding in the middle fusiform gyrus of two brain-damaged prosopagnosic patients. *Neuropsychologia* 47. 2584–2592.
- Stoerig, P. & Cowey, A. (1997). Blindsight in man and monkey. *Brain* 120. 535–559.
- Sur, M., Garraghty, P. E., & Roe, A. W. (1988). Experimentally induced visual projections into auditory thalamus and cortex. *Science* 242(4884). 1437–1441.
- Schwering, A., Krumnack, U., Kühnberger, K.-U., & Gust, H. (2009). Syntactic Principles of Heuristic-Driven Theory Projection. *Cognitive Systems Research* 10(3). 251–269.



- Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience* 19. 109–139.
- Thagard, P. (2008). Cognitive Science. In: Edward N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy*.
- Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Watanabe, S. & Huber, L. (2006). Animal logics: decisions in the absence of human language. *Animal Cognition* 9(4). 235–245.
- Weiller, D., Läer, L., Engel, A. K., & König, P. (2010). Unsupervised learning of reflexive and action-based affordances to model adaptive navigational behavior. *Front Neurobotics* 4. 2.
- Weiskrantz, L. (1986). *A Case Study and Implications*. Oxford: Oxford University Press.
- Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University.
- Wiskott, L. & Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation* 14(4). 715–770.
- Wyss, R., König, P., & Verschure, P. F. M. J. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLOS Biology* 4. e120.
- Zeki, S. (1990). A century of cerebral achromatopsia. *Brain* 113(6). 1721–1777.
- Zihl, J., Cramon, D. von, & Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain* 106. 313–340.
- Zwaan, R., Stanfield, R., & Yaxley, R. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science* 13(2).

**Prof. Dr. Peter König**

University of Osnabrück  
Institute of Cognitive Science  
Albrechtstraße 28  
49069 Osnabrück  
Germany  
pkoenig@uos.de

**Prof. Dr. Kai-Uwe Kühnberger**

University of Osnabrück  
Institute of Cognitive Science  
Albrechtstraße 28  
49069 Osnabrück  
Germany  
kkuehnbe@uos.de

**Tim C. Kietzmann, M. Sc.**

University of Osnabrück  
Institute of Cognitive Science  
Albrechtstraße 28  
49069 Osnabrück  
Germany  
tkietzma@uos.de



Markus Werning, Michela C. Tacca,  
and Aleksandra Mroczko-Wąsowicz

# High- vs Low-Level Cognition and the Neuro-Emulative Theory of Mental Representation

Commentary on Peter König, Kai-Uwe Kühnberger,  
and Tim C. Kietzmann

## 1

König et al. (this volume) in conjunction with König and Krüger (2006) analyze a long-standing and unresolved issue in cognitive science: The relation between low- and high-level cognition. Low cognitive processes include the stages of different perceptual modalities, whereas high-level processes include planning, reasoning, believing, and so on. From the neurophysiological point of view, those systems are based on different networks that correspond to functionally defined cortical regions. However, it seems undeniable that lower perceptual and higher cognitive systems interact. A typical example of a process that occurs at a perceptual stage and communicates with high-level systems is object recognition. During this process, the visual system (i) perceives invariants, i.e., it represents an object as being the same even if perceived from different points of view; and (ii) it subsumes objects that share similar features, for example, different dog instances under the same category: DOG. This representation at the visual level serves as the basis for further higher cognitive processes. From the fact that one recognizes a particular animal as a dog, one can infer a multitude of other facts and events: for example, that this particular dog is not a dangerous one.

The interaction between low- and high-level cognition calls upon the question on whether those systems share similar processes and structures. König and colleagues' core argument is that low- and high-level cognitive systems implement similar structures despite their functional differences. According to the model, the similarity of structure relates to the statistics of the received inputs and the strong relations of perceptual and cognitive systems to action. This approach combines the following assumptions: Neurons in visual areas have sparse activations and are feature specific. They compute the slow temporary changes of an object in order to represent its identity over time. The model purports to explain object representation on the basis of the stability of the input. However, as König and colleagues notice, when subjects represent an object, they often represent this object and its related affordance. An affordance is a property of an object

that allows a subject to perform an action upon that object in a way specific for the object (Gibson, 1977). For example, the perceived affordance of a chair may be ‘sitability’. Affordances may differ depending on the perceiving subject or situation: The same chair that is suitable to sit for a human adult may be perceived as ‘climbable’ by a child or by an adult in a different context (e.g., when using a chair to reach an object). These differences may account for the fact that different individuals across and within different species perceive the world in different ways. Moreover, the representation of affordances is at the root of the ability of an individual to predict and support the sensory consequences of actions. For example, regardless which type of ball is in front of you, you will know that every time you kick a ball, that ball will move in a specific direction. The sameness of behavior when faced with similar stimuli can be described in terms of how the subject learns to generalize predicting the action’s consequences and thereby learns to select an appropriate action.

According to König and colleagues, the generalization occurring during object recognition – subsuming different instances of the same object under the same category and predicting the consequences of our actions over all instances of the same category – is a process similar to the drawing of an analogy. Analogy is a well-studied high-cognitive phenomenon defined as the transfer of knowledge from known examples to unseen examples in order to make the right classification. For instance, an analogy commonly used in science is to compare electrical circuits to hydraulic systems. Sensory systems, like the visual system, might draw inferences from a known scenario to an unknown one in a similar way. In fact, the generalization of invariants and performed actions allows the subject to predict the behavior and perceive the function of newly perceived objects if those objects resemble some of the objects that have already been categorized. Further, it is claimed that a similar mechanism might be at the basis of very simple analogies and that higher cognitive analogies may have a sensory basis. According to the discussed approach, in the course of perceptual analogy representations within the same category will activate similar neuronal groups. For, the objects of a group – e.g., the objects falling under the category BALL – share some aspects. However, those representations also differ to a certain extent between one another. This might depend on the context of the action and the perceived affordance of an object.

Various approaches on the link between low- and high-level cognition highlight the sensory basis of higher-cognitive representations (for a review, see Barsalou, 2008). The pivotal question is what perceptual and higher-cognitive systems share. König and colleagues stress their similarities in terms of representational resources and their structure. We argue that their model based on objec-

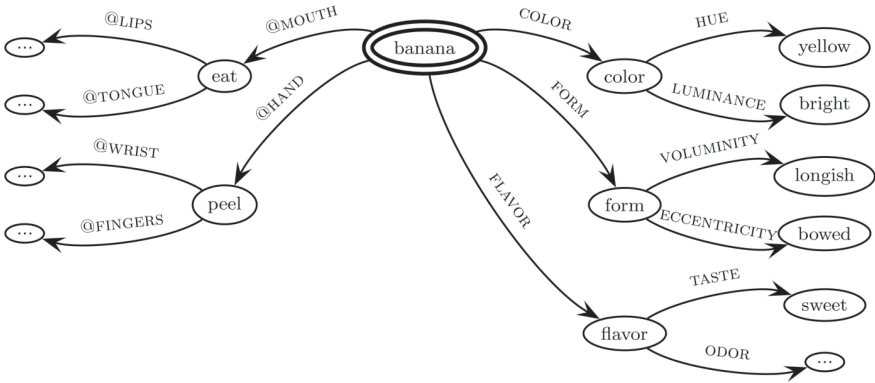
tive function and the representation of affordances explains something more: It also accounts for the distinction between attributive and substance concepts.

## 2

The difference between lower-level sensory representations – prior to object recognition – and higher-level cognitive representations – presupposing object recognition – can be accounted for in terms of the difference between the use of thick substance concepts at the later stage of perceptual representations and that of thin attributive concepts at earlier stages. On the one hand, substance concepts represent stable features, invariant over time, which are governed by the conditions of object identity. For example, a banana no longer falls under the substance concept BANANA when it has been smashed. On the other hand, attributive concepts represent variable features, in the sense that an object can fall under different attribute concepts at different times. For example, an object can have different colors at different times. Substance concepts are typically expressed by concrete nouns – in English by names of individuals like *mama*, names of kinds like *mouse* and names of stuffs like *milk*. Attributive concepts, in contrast, are typically expressed in English by adjectives or abstract nouns: *blue(-ness)*, *warm(-th)*, *lucid(-ity)*. (Millikan, 1998; Werning, 2008, 2010)

The perspective to be developed here largely draws on the theory of neuro-frames (Werning and Maye, 2007). The theory of neuro-frames holds that (i) substance concepts are decomposable into less complex concepts with attributive concepts at the lower levels, that (ii) the compositional structure of a substance concept can be rendered by a recursive attribute-value structure, that (iii) the neural realization of a substance concept is distributed over assemblies of neurons and meta-assemblies thereof, that (iv) those neurons pertain to neural maps for various attributes in many afferent and efferent regions of the cortex, and that (v) object-relative neural synchronization is an appropriate mechanism for binding together the distributed information into the neural realization of the substance concept.

Frame theory provides us with a universal account not only for categorization and its link to action-control, but also for the decomposition of concepts. Frames are recursive attribute-value structures. Attributes assign unique values to objects and thus describe functional relations. The values can be structured frames themselves. A frame is defined for a large domain of things and contains a fixed set of attributes (e.g., color, form, flavor), each of which allows for a number of different values (red, green, etc.). The attributes in question are not constrained



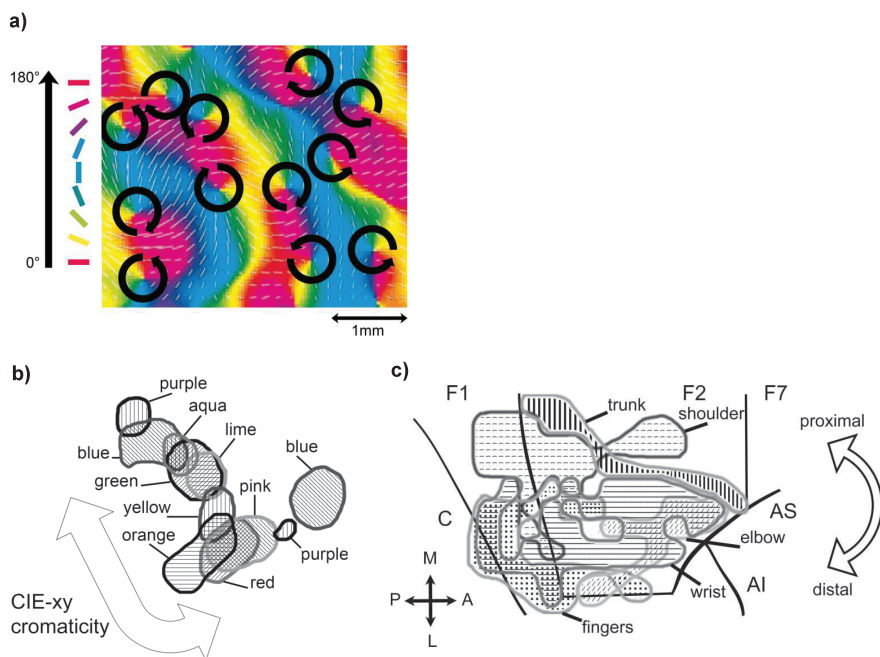
**Figure 1:** Hypothetical fragment of the frame for the concept BANANA. The substance concept to be decomposed is marked by a double-circle as the referring node of the frame. The labeled arrows denote attributes, the nodes their values. Nodes are themselves regarded as concepts and thus as conceptual parts of the central concept. In English, feature attributes (shown on the right) are frequently lexicalized – their arguments typically enter possessive constructions like *The color of the banana is yellow* or *The banana has the color yellow*. Based on linguistic and neurobiological evidence, we assume that affordances often relate to body parts and hence use the convention “@ + body part”. Formally, attributes are mappings from domains of some type into domains of some other type. Petersen and Werning (2007) provide an explicit account of frames using a calculus of typed feature hierarchies and incorporating typicality effects.

to perceptual modalities, but may involve attributes of motor affordances as well. Frames can be nested hierarchically and mutual constraints between attributes (e.g. between states of an object and actions directed to it) and between larger frames can be incorporated. Our model postulates neuro-frames as neuronal bases for concepts.

For many attributes involved in perceptual processing one can anatomically identify cortical correlates. Those areas often exhibit a twofold topological structure and justify the notion of a feature map: (i) a receptor topology (e.g., retinotopy in vision, somatotopy in touch): neighboring regions of neurons code for neighboring regions of the receptor; and (ii) a feature topology: neighboring regions of neurons code for similar features. With respect to the monkey, more than 30 cortical areas forming feature maps are experimentally known for vision alone (Felleman and van Essen, 1991).

Motor attributes may also be parts of frames and appear to have cortical correlates, predominantly in the premotor and motor cortex (Werning, 2010). The cortical organization of motor control with regard to the effectors follows similar topological principles as the cortical organization in perception with regard to the receptors. The discovery of the so-called canonical motor neurons in the mirror

neuron system, activated by the sight of an object to which a certain action is applicable (Rizzolatti and Luppino, 2001; Rizzolatti and Craighero, 2004), may provide a basis to integrate affordances into frames. Figure 2 shows a number of neural maps that relate to various attributes of frames.



**Figure 2:** Cortical realizations of frame attributes.

a) Fragment (ca. 4 mm<sup>2</sup>) of the neural feature map for the attribute orientation of cat V1 (adapted from Shmuel and Grinvald, 2000). The arrows indicate the polar topology of the orientation values represented within each hypercolumn. Hypercolumns are arranged in a retinotopic topology.

b) Color band (ca. 1 mm<sup>2</sup>) from the thin stripes of macaque V2 (adapted from Xiao et al., 2003). The values of the attribute color are arranged in a topology that follows the similarity of hue as defined by the Commission Internationale de l'Eclairages (xy-chromaticity). The topology among the various color bands of V2 is retinotopic.

c) Neural map (ca. 250 mm<sup>2</sup>) of forelimb movement in macaque primary motor (F1) and dorsal premotor cortex (F2, F7) (adapted from Raos et al., 2003). The overarching topology is somatotopic from proximal to distal movement as shown by the arrow. Due to the size of the region one may expect it to comprise maps for more specific motor attributes. C: central sulcus, AS and AI: superior, respectively, inferior arcuate sulcus.

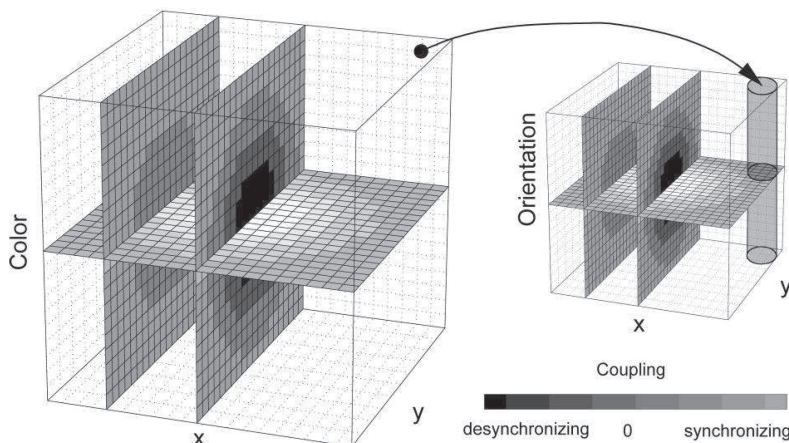
Canonical neurons are involved in mechanisms for recognizing object affordances and carrying out the semantic knowledge about the object (Sahin and Erdogan, 2009). Hence, the activation of the mirror system brings its multimodal neurons to respond not only to action performance, but also to visual, auditory, somatosensory and proprioceptive signals. This suggests that related processes are grounded functionally by multimodal circuits (Gallese and Lakoff, 2005; Rizzolatti and Sinigaglia, 2010). In particular, the intraparietal sulcus and inferior parietal lobule are involved in multisensory integration and vicarious sensory-motor activations (Rizzolatti and Sinigaglia, 2010; Ishida et al., 2010; Rozzi et al., 2006; Bremmer et al., 2001). These regions, able to receive visual input, are directly connected with each other and with the somatosensory cortex (i.e., BA2; Lewis and van Essen, 2000; Pons and Kaas, 1986) integrating tactile and proprioceptive stimuli as well as containing shared sensory-motor representations (Keysers et al., 2010). These multimodal circuits exhibit some basic semantic features. The activation of a specific action concept, e.g. expressing an affordance or any other motor attribute, induces the activation of the multimodal neural circuits (Pulvermüller and Fadiga, 2010).

The fact that values of different attributes may be instantiated by the same object, but are processed in distinct regions of cortex is a version of the binding problem: how is this information integrated in an object-specific way? How can the color and taste of a banana be represented in distinct regions of cortex, although they are part of the representation of one and the same object?

A prominent and experimentally well supported solution postulates oscillatory neural synchronization as a mechanism of binding: Clusters of neurons that are indicative of different properties sometimes show synchronous oscillatory activity, but only when the properties indicated are instantiated by the same object in the perceptual field; otherwise they are firing asynchronously. Synchronous oscillation, thus, might be regarded as fulfilling the task of binding various property representations together to form the representation of an object having these properties (Singer, 1999). Using oscillatory networks as biologically motivated models, it could be demonstrated how the topological organization of information in the cortex by mechanisms of synchronization may yield a logically structured semantics of concepts (Werning and Maye, 2007; Maye and Werning, 2004, see figures 3 and 4). Compositionality theorems have been provided (Werning, 2005). Oscillation functions play the role of object concepts. Clusters of feature sensitive neurons play the role of attributive concepts. The experimental findings by Schnitzler et al. (2006) on the essential role of neural synchronization for action control may justify the extension of the synchrony-based neuro-frame approach from features to affordances. It should be noted that the envisaged semantics is one of emulation: the neuronal structure is partially iso-



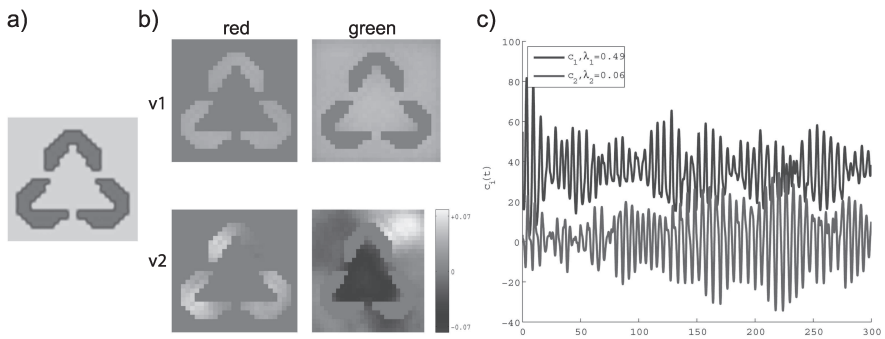
morphic to a (model-theoretic) model of the representational content. A concept like BANANA thus interrelates a.o. sensoric and motoric emulations: Having the concept BANANA means being able to emulate what a banana would look, taste, feel, and smell like and being able to emulate actions afforded by a banana. Triggering the concept activates the respective sensoric and motoric cerebral regions for the purpose of emulation even in the absence of a real banana. The neuro-frame captures how the various sensoric and motoric emulations are linked to each other. Emulative semantics is a non-symbolic, embodied, but still compositional semantics and might be used to link conceptual resources employed in perception and motor planning to linguistic meaning (Werning, 2012).



**Figure 3:** Oscillatory network. The network topology reflects the receptor topology (xy-plane) and the feature topology (z-axis) of the neural maps. Each module realizes one attribute. The layers in each module realize the attribute values. Oscillators activated by neighboring stimulus elements with similar attribute values synchronize (light gray). Oscillators activated by neighboring stimulus elements with unlike attribute values de-synchronize (dark gray). The layers of different modules are connected in a synchronizing way that respects the common receptor topology. (From Maye and Werning, 2007).

Support for the theory of neuro-frames also comes from a number of neuro-linguistic studies. Based on a review of neurobiological data, Pulvermüller (1999) suggests that neural assemblies that pertain to the sensory-motor cortices and are bound by neural synchronization play an important role in understanding the meanings of words and sentences. These cortical sensory-motor action and perception circuits are interdependent in language comprehension. Neuroim-

aging investigations have shown that perception and understanding of stimuli depend on motor circuits, i.e. specific motor activations can be found when subjects understand speech sounds, word meanings, semantic categories and sentence structures (Pulvermüller and Fadiga, 2010). FMRI studies (Pulvermüller, 2005) regarding the understanding of verbs, e.g., hint at a differential top-down activation of motor and pre-motors areas. We know that the understanding of concrete nouns like *hammer*, for which not only features, but also affordances are salient, results in an activity distributed over the premotor and the visual cortex (Martin et al., 1996; Martin, 2007). The hypothesis that words for substance concepts arouse more widely distributed activity than words for attributive concepts has also been supported by EEG studies (Rappelsberger et al., 2000). Brain areas involved in motor control contribute to neural networks in which verb representations are grounded, e.g. studies on motor deficits such as Parkinson disease reveal impairment of patients' action naming (Rodríguez-Ferreiro et al., 2009). Higher-order abilities such as thinking or linguistic concept use are based in sensory-motor abilities. The relation to attentional mechanisms has been studied by Tacca (2010). Parallels to cases of synaesthesia where hyperbinding within neuroframes might play a role have been discussed by Mroczko-Wąsowicz and Werning (2012).



**Figure 4:** An oscillator network with a single module for color with layers for red and green is stimulated with the Kanizsa illusion. a) Stimulus: three red circle segments on a green ground. b) The two strongest eigenmodes of the network dynamics  $v_1$  and  $v_2$ , each subdivided according to layers, are shown. The signs of the vector components are indicated by shades of gray: light gray: positive, middle gray: zero, dark gray: negative. c) Temporal evolution of the two eigenmodes are given by the characteristic oscillatory functions  $c_1(t)$  and  $c_2(t)$ . The eigenvalues  $\lambda_{1,2}$  yield the relative contribution of each eigenmode to the overall variability of the network dynamics. Semantic interpretation: The first eigenmode does not render figure ground segregation. The second eigenmode, however, renders a representation of the illusory triangle (object concept:  $-c_2$ ) as distinct from the background (mostly zero) and the united circle segments (object concept:  $+c_2$ ).

### 3 Conclusions

We argue that König and colleagues' model, which highlights the optimization of feature selectivity and feature predictability, may also contribute to the explanation of a further property of high- and low- level processes: The distinction between substance and attributive concepts. Neuroframe theory gives a detailed account of how substance concepts – presupposed for higher cognitive processes – and attributive concepts – hosted by lower perceptual cortical areas – relate to each other. The relation is one of recursive conceptual decomposition. Due to the interaction between affordance and feature attributes, neuroframes are flexible enough to allow for a situational dependency when it comes to feature selection. While attributive concepts specify volatile properties of objects, substance concepts are governed by the identity conditions of objects and thus warrant a stable identification of those objects. Since neuroframes capture how substance and attributive concepts relate to each other a situation dependent way they enable an optimization of predictability.

A main idea in philosophy is that if perception and cognition interact, they need to have the same type of representational content, or if they do have different types of content, one needs to further explain how their contents relate. Our hypothesis is that representations at the cognitive level involve conceptual representations (substance concepts) that derive from the recombination of primitive attribute concepts that occur at earlier stages. König and colleagues instead argue that the differences in the final make-up of the representation between low- and high-level cognition account for those systems to implement distinct kinds of content. A hallmark of conceptuality is that representations combine in a compositional fashion. As we noticed above, the recombination of attributive perceptual and motor representations into substance concepts satisfies the principle of compositionality. Hence, we argue that those representations have conceptual content, even if they are not symbolic representations, but emulations.

### References

- Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In: Kittay, E. & Lehrer, A. (eds.). *Frames, fields, and contrasts: New essays in semantic and lexical organization*. Hillsdale, NJ: Lawrence Erlbaum Associates. 21–74.
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology* 59. 617–645.
- Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.
- Bremmer, F., Schlack, A., Shah, N. J., Zafiris, O., Kubischik, M., Hoffmann, K.-P., Zilles, K., & Fink, G. R. (2001). Polymodal motion processing in posterior parietal and premotor cortex:

- a human fMRI study strongly implies equivalencies between humans and monkeys. *Neuron* 29. 287–296.
- Felleman, D. J. & van Essen, D. C. (2003). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1. 1–47.
- Fodor, J. & Lepore, E. (1992). *Holism: A shopper's guide*. Oxford: Blackwell.
- Gallese V. & Lakoff, G. (2005). The Brain's Concepts: The Role of the Sensory-Motor System in Reason and Language. *Cognitive Neuropsychology* 22. 455–479.
- Gibson, J. J. (1977). The theory of affordances. *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. Hillsdale, NJ: Lawrence Erlbaum. 67–82.
- Goldstone, R. & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition* 65. 231–262.
- Ishida, H., Nakajima, K., Inase, M., & Murata, A. (2010). Shared mapping of own and others' bodies in visuotactile bimodal area of monkey parietal cortex. *Journal of Cognitive Neuroscience* 22. 83–96.
- Keysers, C., Kaas, J. H., & Gazzola, V. (2010). Somatosensation in social perception. *Nature Reviews Neuroscience* 11. 417–28.
- König, P. & Krüger, N. (2006). Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics* 94(4). 325–334.
- Lewis, J. W. & Essen, D. C. van (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *Journal of Comparative Neurology* 428. 112–137.
- Malsburg, C. von der (1981). *The correlation theory of brain function (Internal Report No. 81–2)*. Göttingen: MPI for Biophysical Chemistry.
- Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology* 58(1). 25–45.
- Martin, A., Wiggs, C. L., Ungerleider, L., & Haxby, J. V. (1996). Neural correlates of category specific knowledge. *Nature* 379. 649–652.
- Maye, A. & Werning, M. (2004). Temporal binding of non-uniform objects. *Neurocomputing* 58–60. 941–948.
- Maye, A. & Werning, M. (2007). Neuronal synchronization: From dynamic feature binding to object representations. *Chaos and Complexity Letters* 2. 315–325.
- Millikan, R. G. (1998). A Common Structure for Concepts of Individuals, Stuffs and Real Kinds: More Mama, More Milk, and More Mouse. *Behavioral and Brain Sciences* 21. 55–100.
- Mroczko-Wąsowicz, A. & Werning, M. (2012). Synesthesia, sensory-motor contingency and semantic emulation: How swimming style-color synesthesia challenges the traditional view of synesthesia. *Frontiers in Psychology* 3(279). 1–12.
- Petersen, W. & Werning, M. (2007). Conceptual fingerprints: Lexical decomposition by means of frames – a neuro-cognitive model. In: Priss, U., Polovina, S., & Hill, R. (eds.), *Conceptual structures: Knowledge architectures for smart applications* LNAI 4604. 415–428.
- Pons, T. P. & Kaas, J. H. (1986). Corticocortical connections of area 2 of somatosensory cortex in macaque monkeys: a correlative anatomical and electrophysiological study. *Journal of Comparative Neurology* 248. 313–335.
- Prinz, J. J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*. Cambridge, MA: MIT Press.
- Pulvermüller, F. (1999). Words in the Brain's Language. *Behavioral and Brain Sciences* 22. 253–279.

- Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience* 6(7). 576–582.
- Pulvermüller, F. & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience* 11(5). 351–360.
- Pulvermüller, F., Lutzenberger, W., & Preissl, H. (1999). Nouns and Verbs in the Intact Brain: Evidence from Event-related Potentials and Highfrequency Cortical Responses. *Cerebral Cortex* 9(5). 497–506.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: MIT Press.
- Raos, V., Franchi, G., Gallese, V., & Fogassi, L. (2003). Somatotopic Organization of the Lateral Part of Area F2 (Dorsal Premotor Cortex) of the Macaque Monkey. *Journal of Neurophysiology* 89. 1503–1518.
- Rappelsberger, P., Weiss, S., & Schack, B. (2000). Coherence and phase relations between EEG traces recorded from different locations. In: Miller, R. (ed.). *Time and the brain*. Amsterdam: Harwood Academic Publishers. 297–330.
- Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience* 27. 169–192.
- Rizzolatti, G. & Luppino, G. (2001). The cortical motor system. *Neuron* 31. 889–901.
- Rizzolatti, G. & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Reviews Neuroscience* 11(4). 264–274.
- Rodríguez-Ferreiro, J., Menéndez, M., Ribacoba, R., & Cuetos, F. (2009). Action naming is impaired in Parkinson disease patients. *Neuropsychologia* 47. 3271–3274.
- Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology* 8. 382–439.
- Rozzi, S., Calzavara, R., Belmalih, A., Borra, E., Gregoriou, G. G., Matelli, M., & Luppino, G. (2006). Cortical connections of the inferior parietal cortical convexity of the macaque monkey. *Cerebral Cortex* 16. 1389–1417.
- Sahin, E. & Erdogan, S. T. (2009). Towards linking affordances with mirror/canonical neurons. *ISCI*. 397–404.
- Schnitzler, A., Timmermann, L., & Gross, J. (2006). Physiological and pathological oscillatory networks in the human motor system. *Journal of Physiology, Neuronal Dynamics and Cortical Oscillations* 99. 3–7.
- Shmuel, A. & Grinvald, A. (2000). Coexistence of linear zones and pinwheels within orientation maps in cat visual cortex. *Proceedings of the National Academy of Sciences USA* 97. 5568–5573.
- Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron* 24. 49–65.
- Tacca, M. C. (2010). *Seeing objects: The structure of visual representation*. Paderborn: Mentis.
- Treisman, A. (1993). The perception of features and objects. In: Baddeley, A. & Weiskrantz, L. (eds.). *Attention: Selection, awareness and control*. Oxford: Oxford University Press.
- Werning, M. (2005). The temporal dimension of thought: Cortical foundations of predicative representation. *Synthese* 146(1/2). 203–24.
- Werning, M. & Maye, A. (2005). Frames, coherency chains and hierarchical binding: The cortical implementation of complex concepts. In: Bara, B. G., Barsalou, L., & Bucciarelli, M. (eds.). *Proceedings of the twenty-seventh annual Conference of the Cognitive Science Society*. New York: Erlbaum. 2347–2352.

- Werning, M. & Maye, A. (2007). The cortical implementation of complex attribute and substance concepts: Synchrony, frames, and hierarchical binding. *Chaos and Complexity Letters* 2(2/3). 435–452.
- Werning, M. (2008). The complex first paradox – Why do semantically thick concepts so early lexicalize as nouns? *Interaction Studies* 9(1). 67–83.
- Werning, M. (2010). Complex first? On the evolutionary and developmental priority of semantically thick words. *Philosophy of Science* 77. 1096–1108.
- Werning, M. (2012). Non-symbolic compositional representation and its neuronal foundation: Towards an emulative semantics. In: Werning, M., Hinzen, W., & Machery, M. (eds.). *The Oxford Handbook of Compositionality*. Oxford: Oxford University Press. 633–654.
- Xiao, Y., Wang, Y., & Felleman, D. J. (2003). A spatially organized representation of colour in macaque cortical area V2. *Nature* 421. 535–539.

**Prof. Dr. Markus Werning**

Professor of Philosophy of Language and Cognition  
Ruhr-University Bochum  
Department of Philosophy  
44780 Bochum  
Germany  
markus.werning@ruhr-uni-bochum.de

**Dr. Michela C. Tacca**

Chair of Theoretical Philosophy  
Heinrich-Heine-University Düsseldorf  
Department of Philosophy  
Universitätsstr. 1  
40225 Düsseldorf  
Germany  
tacca@phil-fak.uni-duesseldorf.de

**Aleksandra Mroczko-Wąsowicz, PhD**

Assistant Professor  
National Yang-Ming University Taipei  
Institute of Philosophy of Mind and Cognition  
155, Sec. 2, LiNong St., Beitou  
Taipei 11221  
Taiwan (R.O.C.)  
mroczko-wasowicz@hotmail.com

Reinhold Kliegl and Ralf Engbert

# Evaluating a Computational Model of Eye-Movement Control in Reading

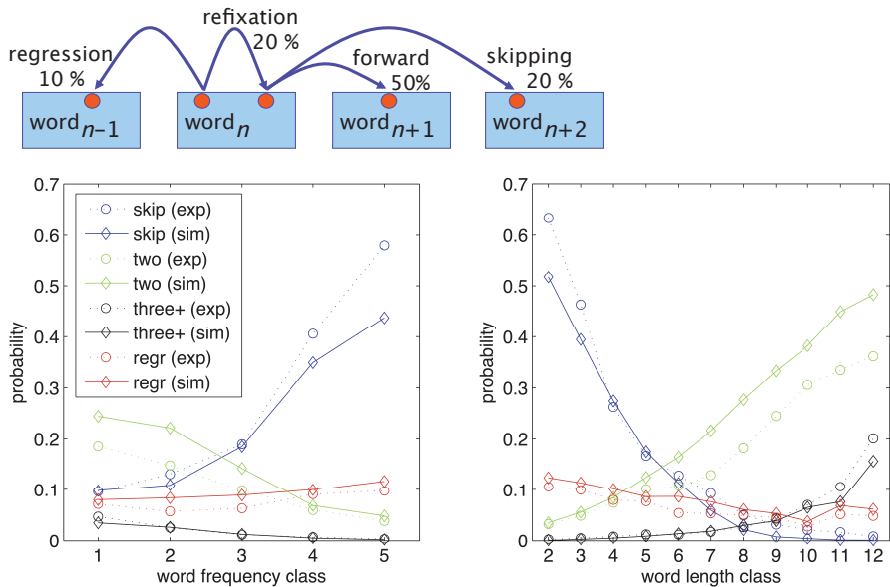
## 1 Some basic facts about eye movements during reading

Reading is an activity we all engage in on a daily basis. It requires the coordination of perceptual and oculomotor processes as well as the integration of this new information provided by perception and eye movements with the available knowledge and expectations about what is being read. This chapter is about how we evaluate a model about an important component of this complex process, that is a model about how we move our eyes across the words of a sentence. We present this model as a prototype of an integrated theoretical, computational, and data-analytic approach for the interface between experimental psychology, cognitive (neuro-)science and computational neuroscience. With examples from ongoing research we illustrate how the model can be evaluated against a set of criteria for strong model tests, comprising goodness of fit, strictness of model, reliability of data, and unexpected predictions (Roberts and Pashler, 2000). Before we turn to these model tests we introduce some basic facts about eye movements during reading and (a subset of) the theoretical principles that guided the construction of our model.

When we record or directly observe a reader's eye movements, it is immediately apparent that the eyes do not move continuously across the words of a sentence.<sup>1</sup> Counter to what introspection suggests that we do most of the time; we notice a strict alternation of quick movements, called saccades (lasting about 20 to 30 ms), and periods of relative rest, called fixations (with mean durations ranging from 150 to 350 ms). Visual input about what we read occurs only during fixations; we are effectively blind while the eyes are in flight and indeed most of the time we are not aware of saccades. Thus, what we experience as a continuous movement across the words of a sentence is largely a construction of the mind.

---

<sup>1</sup> The results reported in this chapter are based on binocular measurements at 250 Hz or 500 Hz from 273 readers who read 144 isolated sentences (i.e., the Potsdam Sentence Corpus; Kliegl, Grabner, Rolfs, and Engbert, 2004; Kliegl, Nuthmann, and Engbert, 2006). We map fixation positions to a specific letter within a word. The fixation positions differ slightly between the two eyes. Our results are based on measurements of the right eye.



**Figure 1:** (top) Illustration of four types of eye movements and their marginal probabilities during normal reading of German prose. Experimental (exp) and simulated (sim) values for skipping (skip), regression (regr), and fixation probabilities (2 or  $\geq 3$ ) conditional on word length (bottom left) and on log10 word frequency per million words (bottom right) (modified from Engbert et al., 2005).

The discrete nature of fixation-saccade cycles suggests a simple taxonomy of eye movements relative to the words of the sentence. As illustrated in Figure 1 (top panel), we distinguish four types: roughly 50% of the saccades carry the eyes from one word to the next word, about 20% of them shift the position within the currently fixated word, about 20% skip the next word, and about 10% of the time we move back to an earlier word. These statistics greatly depend on word properties, most notably on the lengths of the words. As shown in Figure 1 (bottom left), skipping probability decreases strongly from roughly 60% for two-letter words to less than 1% for 12-letter words; conversely the probability of refixations increases from close to 0% to around 40%. The same statistics can also be plotted over the frequency with which words are observed. As shown in Figure 1 (bottom right), skipping and refixation probability decreases and increases with log-frequency of observing a word in texts comprising one million words.

Fixation probabilities yield one key set of dependent measures. The second set of measures relates to different types of fixation durations. Here we distinguish, for example, between durations of fixations when they are the only fixa-



tion on a word (i.e., a single fixation duration), the first of two fixations, or the second of two fixations. Usually, we also sum all the fixations on a word to a measure of total reading time. Again, all these measures exhibit systematic relations to the lengths and frequencies of words on which they are observed: The longer or the less frequent a word, the longer the fixation duration or total fixation time.

The length and type frequency of the fixated word are but two highly correlated properties of a large number of variables that have been shown to influence fixation probabilities and durations (e.g., Rayner, 1998, for a review). Other variables are, for example, the predictability of a word given the prior words of the sentence. This measure is usually obtained in independent studies in which subjects have to guess the words of a sentence in an incremental order. Other variables reflect how similar the word is to other words of the language, measured by how many words can be derived from a word if one allows to exchange one letter (i.e., an edit distance of 1). Another example is the informativeness of the beginning of a word for its identification. For example, given a long word with “xy...” as initial letters, there are not many alternatives to “xylophone”. Moreover and critically, properties of words  $n-1$  or  $n+1$  have been shown to influence fixation durations on word  $n$  (e.g., Kliegl et al., 2006). We will describe a few of these and a few additional effects in the context of introducing some of the theoretical principles guiding the evaluation of our computational model. In summary, despite the one-dimensional space during reading a single sentence on a line, the associated eye movements exhibit a very complex trajectory modulated by a large number of variables.

## 2 Theoretical principles of eye-movement control during reading

Statistics of various types of fixation probabilities and fixation durations serve as benchmark data for all computational models of eye movements during reading. Usually, for each model a number of free parameters is estimated such that summary statistics as those described in Figure 1 are reproduced. The bottom panels of Figure 1 show that our model, called SWIFT, does a good job of recovering fixation probabilities; it does not show that the model also accounts for the differentiated pattern of fixation durations and distributions of landing positions in words contingent on word length and the amplitude of the last saccade (Engbert,

Nuthmann, Richter, and Kliegl, 2005).<sup>2</sup> The model is not unique in this respect – there are at least three other models that can account for such data (McDonald, Carpenter, B., and Shillcock, 2006; Reichle, Pollatsek, Fisher, and Rayner, 1998; Reilly and Radach, 2006; for a comprehensive review of these and other models see Reichle, Pollatsek, and Rayner, 2003). Given this state of the research we can describe how we plan to compare such models in a principled way. But before we turn to the issue of model comparison, evaluation, and development, we describe two of seven core principles that guided the implementation of the SWIFT model: (1) the distinction between when and where to move the eyes and (2) the notion of the perceptual span. These theoretical principles have always been formulated in a qualitative way. Only their implementation in a computer program requires a commitment to a specific mathematical representation.

## 2.1 When and where to move the eyes

When reading these lines, the control of our eye movements is based on principles about *when* and *where* to launch the next saccade that moves the next word into the fovea for high-acuity analysis. Interestingly, neurophysiological evidence suggests that the temporal and spatial aspects of saccade-generation are largely independent across several levels of organization (Findlay and Walker, 1999).

The temporal aspect of eye-movement behavior, the *when* decision, is captured by fixation duration measures. Over the past 30 years, much research has been conducted to determine the relationship between fixation durations and linguistic and/or oculomotor variables. First, as described above, it has been shown that various lexical, syntactic, and discourse factors influence fixation durations on words. Thus, fixation durations in reading are sensitive to local processing difficulty (Rayner, 1998). Second, fixation durations are also influenced by low-level nonlinguistic factors. As a consequence, there are fundamental modulations of fixation durations by word length, within-word fixation position, and the distance between fixation locations (i.e., launch site of last saccade), which are unrelated to word recognition (Vitu, McConkie, Kerr, and O'Regan, 2001).

The *where* pathway, i.e., the question of spatial selection for the next saccade, must solve two tasks: First, which word is to be selected as the target of the next saccade, and, second, what are the principles underlying the control of within-word landing position. The mean value of the landing position distribution is termed the preferred viewing location (Rayner, 1979), which is on average slightly

---

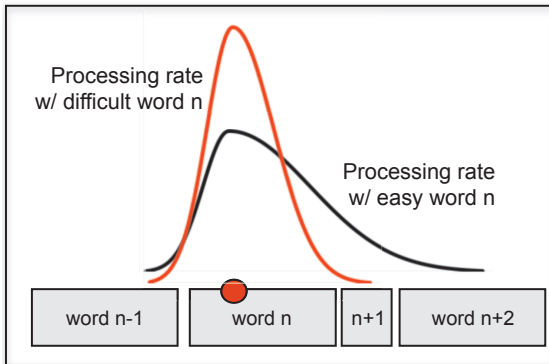
<sup>2</sup> <http://www.agnld.uni-potsdam.de/~ralf/swift/>.

left of the word center. The initial landing position gives rise to the roughly parabolic refixation probability effect, the optimal viewing position (McConkie, Kerr, Reddix, Zola, and Jacobs, 1989), and the inverted-optimal viewing position (IOVP) for fixation durations (Vitu et al., 2001). The assumption that there are partially independent *when* and *where* pathways (Schad and Engbert, 2012) provides an important boundary condition for the development of psychologically plausible theoretical models of eye-movement control during reading.

## 2.2 Perceptual span

Analyses of large corpora of eye movements strongly suggest that non-local (distributed) effects of word difficulty on eye fixations during reading are likely to be much more pervasive than suggested by research examining only a few experimentally manipulated target words per sentence (Kliegl et al., 2006; Kliegl, 2007). Distributed processing means that the fixation duration on a word is not only influenced by the characteristics of the fixated word itself but – due to graded parallel word processing within the perceptual span – depends also on the characteristics of the word to the left (lag effect) as well as those of the word to the right of fixation (successor effect). The perceptual span covers an area roughly extending about 15 characters to the right and three characters to the left of the point of fixation in alphabetic languages (McConkie and Rayner, 1975; Rayner, 1975; see Figure 2 for an illustration). Given the decrease of visual acuity with an increase in the eccentricity of information relative to the fixation position, the rate of processing (represented by the height of the functions in Figure 2) is expected to decline. The asymmetry in the direction of reading is interpreted as a modulation of visual perception by attention. Determining which type of information (i.e., visual, sublexical, lexical, semantic) is available from the upcoming word is an area of active and controversial research (see Kliegl et al., 2006; Kliegl, 2007; Rayner, Pollatsek, Drieghe, Slattery, and Reichle, 2007).

Figure 2 also illustrates the proposal that the perceptual span is dynamically modulated by the difficulty of local processing (Schad and Engbert, 2012). Specifically, the peak of processing rate may be lower and the rightward extension may be larger when the eye rests on highly frequent or predictable words in comparison to fixations on low or unpredictable words. Such a dynamical modulation of the perceptual/attentional span implies that processing of a fixated difficult word is processed at a higher rate than on average, but at the cost of reduced parafoveal processing of the upcoming word (Henderson and Ferreira, 1990; Inhoff and Rayner, 1986; Rayner and Pollatsek, 1987).



**Figure 2:** Perceptual/attentional span for a fixation on word  $n$  (•) and its dynamical modulation by the difficulty of word  $n$ . Height of curves represents processing rates for a difficult and an easy word  $n$ .

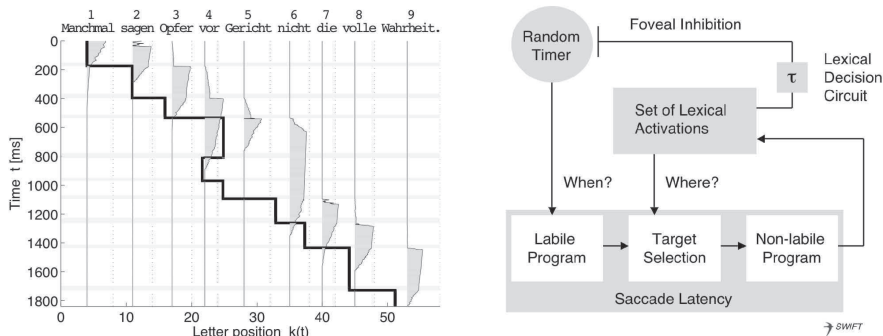
## 2.3 Other core principles

Separate pathways for saccade timing and saccade target selection and spatially distributed processing of an activation field are two of seven core principles of the SWIFT model (Engbert et al., 2005; Table 1). The other five relate to timing of saccade programs and to how this timer can be inhibited with a delay by foveal processing difficulty, to a distinction between different stages in a saccade program, to systematic and random errors in saccade lengths and the implications for mislocated fixations due to such error, and to the relation between saccade latency and saccade amplitude.

## 3 Computational models of eye-movement control in reading

In Figure 3a we illustrate how the SWIFT model simulates reading the sentence “Manchmal sagen Opfer vor Gericht nicht die volle Wahrheit” [Sometimes victims do not say the complete truth in court]. With time running from top to bottom, the black line indicates the position of the eye at a given moment. Thus, vertical black lines represent fixation durations and horizontal black lines indicate saccades.

The grey hills beneath each word in Figure 3a, show the dynamics of the activation field for the nine words of this sentence in this simulation. Note that some



**Figure 3:** The SWIFT model. Left: Example of a numerical simulation of the SWIFT model. Saccade target selection is driven by a spatially-distributed activation field. Word-based activations are illustrated by the shaded areas. The eye's scanpath is indicated by the black line (from Engbert et al., 2005, Fig. 7). Right: Schematic representation of SWIFT. Two independent pathways control fixation duration (“when”) and saccade target selection (“where”). A random timer controlling fixation durations can be inhibited to adjust fixation duration to foveal processing difficulty (from Engbert et al., 2005, Fig. 6).

of the activations overlap in time; there is parallel distributed processing. For example, during the first fixation the first three words are active at one point in time. The rate of processing these words depends on how far the word is from the current fixation position. Activation rises steeply for the first, less steeply for the second, and very slowly for the third word. After the first saccade, the rate rises steeply for the third word; also the fourth word is activated because it is now in the perceptual span.

In general, activations related to the  $N$  words of a sentence are governed by an  $N$ -dimensional set of coupled ordinary differential equations,

$$\frac{d}{dt} a_n(t) = F_n(t) \Lambda_n(t) - \omega \quad (n = 1, 2, 3, \dots, N), \quad (1)$$

where  $\Lambda_n(t)$  is the processing rate,  $F_n(t)$  is a preprocessing factor, which introduces a fast buildup of activation in an early processing stage and is modulated by word predictability  $p_n$ , and  $\omega$  is a global decay process, which we interpret as a memory leakage (see Engbert et al., 2005, for more details of the mathematical formulation of SWIFT).

Turning one such activation “hill” 90° degrees counterclockwise, processing a word means that two random walks spliced at the peak are completed. In SWIFT, word processing difficulty is modulated by printed word frequency and predictability. Low-frequency words have high peaks; high-frequency words have very small peaks. For example, you barely notice the activation associated with

“sich”, a reflexive pronoun. Thus, we assume that word frequency,  $f_n$ , is related to the maximum of activation,  $L_n$ ,

$$L_n = \alpha - \beta \log f_n, \quad (2)$$

while predictability,  $p_n$ , modulates processing rate. High values of predictability decrease processing rate during parafoveal preprocessing and increase processing rate during lexical completion. Given this dynamics, effects of word length are the consequence of an asymmetric Gaussian-type distribution of processing rate around the current fixation position as shown in Figure 2.

The second principle of separate where-and-when pathways specifies that we must distinguish between target selection (Where?) and timing of saccades (When?). As illustrated in Figure 2b for SWIFT, saccade generation begins with a random timer inducing the start of the next saccade program. The probability to select a word as the next saccade target is computed from its relative lexical activation (i.e., the word’s activation value divided by the sum of all lexical activations). Saccade execution occurs only after the necessary saccade-program latency and thereby induces a delay between the effect of lexical activation on target selection and the effect of this saccade on the processing rates in the dynamical field of activations. The set of lexical activations causes also foveal inhibition on the start of the next saccade program (see right panel of Figure 3). Again, this long-loop lexical control process takes time, and this second type of delay is captured within the model parameter  $\tau$ .

Thus, the movement of the eye depends stochastically on lexical activation of the field of words at the time when this decision is made. Those with high activation are more likely to be selected as saccade targets. For example, for the first saccade of Figure 1, the second and the third word have activation above zero and the second word “won” the competition for being selected as saccade target. Importantly, this single principle of target selection generates all types of eye movements introduced in Figure 1: movement to the next word, skipping, refixations, and regressions. None of the competitor models are as “parsimonious” in this respect.

In the initial versions of SWIFT, processing rate for letters was assumed to follow an asymmetric Gaussian distribution with different parameters,  $\sigma_R$  and  $\sigma_L$ , representing the extension of the span to the right and to the left of the fixation point, respectively. Basically, one of the curves shown in Figure 2 was assumed to apply throughout a simulation. In the dynamical systems framework of SWIFT, discrete processing cycles (“sense” → “think” → “act”) are replaced by the temporally continuous evolution of a set of mutually interacting variables representing different cognitive subsystems (Beer, 2000). Within such a framework, it is con-

ceptually possible to implement the dynamic interactions between subsystems very precisely.

We illustrate such a model modification with the proposal of a dynamical span (Schad and Engbert, 2012). When the reading material is difficult, the size of the perceptual span is smaller than for a text of average difficulty (Henderson and Ferreira, 1990). As illustrated in Figure 2, this effect can be accounted for with a sharper distribution of the processing span, determined by parameters  $\sigma_R$  and  $\sigma_L$ , for increasing foveal word difficulty represented by a higher average foveal activation  $a_k(t)$  at time  $t$ . Specifically, Engbert (2007; see also Schad and Engbert, 2012) assumes that (i) the extension of the processing span to the left is constant; and given by parameter  $\sigma_L$ , (ii) the processing span is symmetric for high foveal load  $a_k(t)$ ; and (iii) the extension to the right,  $\sigma_R$ , increases with decreasing foveal load  $a_k(t)$ .

This leads to the following relation:

$$\sigma_R = \sigma_L + \delta_1 F(a_k(t)) \quad (3)$$

where  $\delta_1$  is a free parameter representing the strength of the dynamical control mechanism and  $F(a)$  is a sigmoid function.

This is only the beginning. To explore the viability of the concept of a dynamic perceptual span, different mathematical formulations must be implemented and tested by computer simulations and statistical analyses. For example, in ongoing simulations of data from German-English bilingual readers, the Gaussian-type processing span was not constrained enough by data due to its long tail. Therefore, we revised the functional form of the processing span to an inverted quadratic form. As an important property of such a span, we obtained sharp edges of processing. Based on this modification, we were able to estimate the dynamic part of the processing span.

## 4 Model analysis and comparison

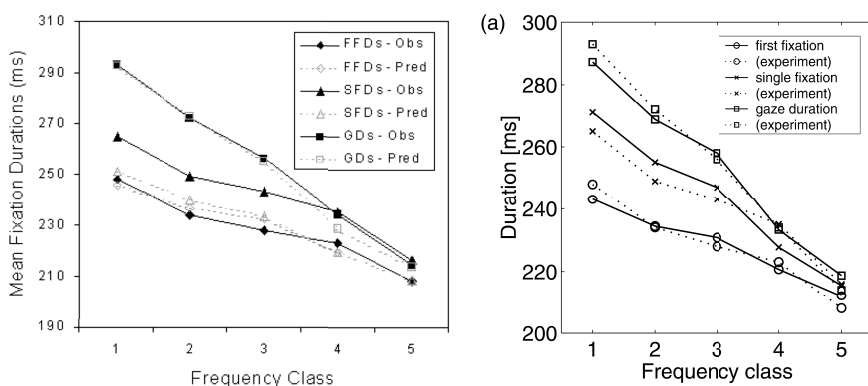
We developed the SWIFT model (Engbert, Longtin, and Kliegl, 2002, Engbert et al., 2005) on the assumption of spatially distributed processing. Another model implementing this assumption is Glenmore (Reilly and Radach, 2003, 2006). In contrast, the E-Z Reader model (Reichle et al., 1998, Reichle, Pollatsek, and Rayner, 2006) is built on the notion of sequential attention shifts (SAS). For the class of SAS models, the serial allocation of visual attention from one word to the next is the central principle driving eye movements. We also contributed to

the SAS-line of research by proposing a model with fewer internal states based on advanced stochastic principles (semi-Markov processes; Engbert and Kliegl, 2001).

Finally, the SERIF model is another fully implemented model which builds upon functional implications of an apparent vertical splitting of the fovea (McDonald et al., 2005).

The development of such quantitative psychological theories is a clear signature of scientific progress. Given the range of competing models of eye-movement control during reading, the need for mathematical analyses and comparisons of models is obvious. Of course, the problem of model selection, analysis, and comparison is a growing area of research in cognitive science in general (e.g., Special Issue on “Model Selection” in the *Journal of Mathematical Psychology*, 2002).

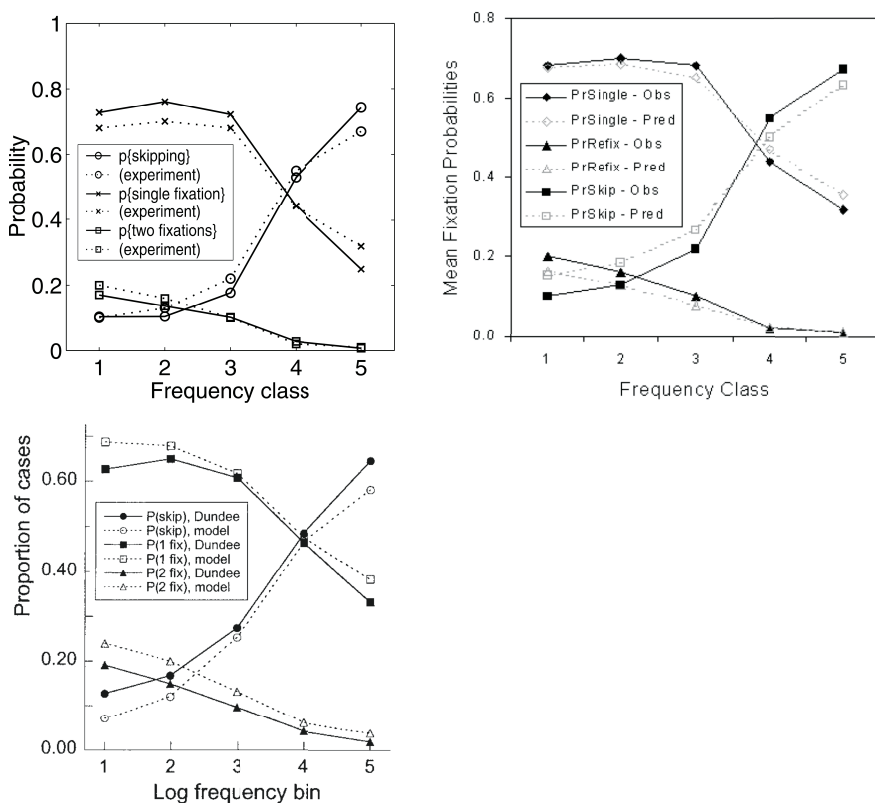
So how should we compare different mathematical/computational models? Model comparisons are typically based on goodness-of-fit (GOF) statistics, which quantify how much a model’s prediction deviates from a given set of experimental data. When using GOF, the underlying assumption is that the model producing the best fit to all data must be a closer approximation to the underlying cognitive process. However, because of random variation in the experimental and statistical methods used (e.g., repeated measurements, inferential statistics), model comparisons based on GOF alone will, in general, produce misleading results (Roberts and Pashler, 2000).



**Figure 4:** Both (a) the initial SWIFT model (Engbert et al., 2002, Fig. 5a) and (b) the E-Z Reader model (Reichle et al., 2003, Fig. 6 top panel) were fit to the same set of fixation durations measured for 5 different word-frequency classes (data from Schilling et al., 1998). Lines refer to observed (Obs) and predicted (Pred) gaze durations (GDs), single fixation durations (SFDs), and first fixation durations (FFDs).



Moreover, advanced mathematical models often fit experimental data equally well. For example, SWIFT (Engbert et al., 2002) and E-Z Reader (Reichle et al., 2006) reproduced fixation durations reported in Schilling, Rayner, and Chumbley (1998) equally well in terms of GOF (see Figure 4). The lines represent observed and predicted gaze durations (i.e., the sum of fixations when a word is first read), single-fixation durations (i.e., when a word is fixated exactly once), and first-fixation durations (i.e., the duration of the first fixation, irrespective of how many fixations occurred) as a function of five log-frequency classes.



**Figure 5:** Data and simulations of single-fixation, skipping, and refixation probabilities as a function of log<sub>10</sub> word frequency (per million words) for the E-Z Reader model (left, Reichle et al., 2003, Fig. 6 bottom panel) and the initial SWIFT model (middle, Engbert et al., 2002, Fig. 5b) were fit to the same set of fixation durations measured for 5 different word-frequency classes (data from English sentences; Schilling et al., 1998). The right panel shows the fit of the SERIF model (McDonald et al., 2005, Fig. 3b; data from English newspaper texts [Dundee corpus], Kennedy, 2003).

Figure 5 (left and middle panel) illustrates the similarity in fit relating to probabilities for single fixations, skipping, or refixations as a function of word frequency for the same data (Schilling et al., 1998) and the same two models (E-Reader 9, Reichle et al., 2006; SWIFT, Engbert et al., 2002). Interestingly, the simulations of the SERIF model (McDonald et al., 2005) led to a very similar pattern of means although the model was fit to a completely different set of English eye movement data (i.e., Dundee Corpus; Kennedy 2003).<sup>3</sup>

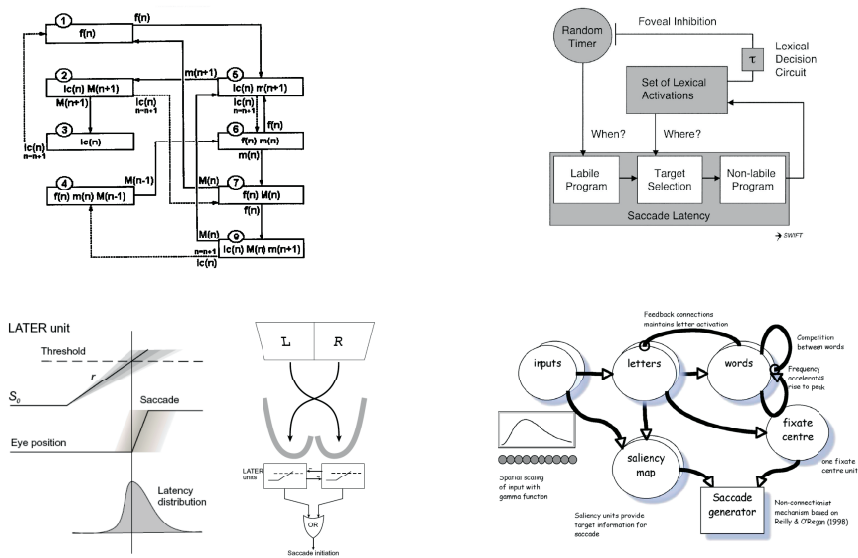
We consider the qualitative (and also largely quantitative) agreement between observations and simulations as quite remarkable. Therefore, to restate the argument from above, for this research field at this point in time, the qualitative agreement between model and data is more important than differences in quantitative goodness-of-fit statistics (such as root mean squared error).

If we grant comparability in goodness of fit, we may still compare the models with respect to their complexity. One indicator of model complexity is the number of free model parameters. Interestingly, the models do not differ on this dimension either, ranging from 13 to 18. Moreover, many of the parameters are only free to vary within a range dictated by substantive issues. For example, in the SWIFT model, the asymmetry of the span, estimated as Gaussian standard deviations for the left and right processing-rate functions (see Figure 2), must map onto a plausible range of letters (i.e., 3 to the left and around 10 to the right). Similarly, parameters estimating delay lines in the model are narrowly constrained by the physiology of the structures known to be involved in saccade programming and execution.

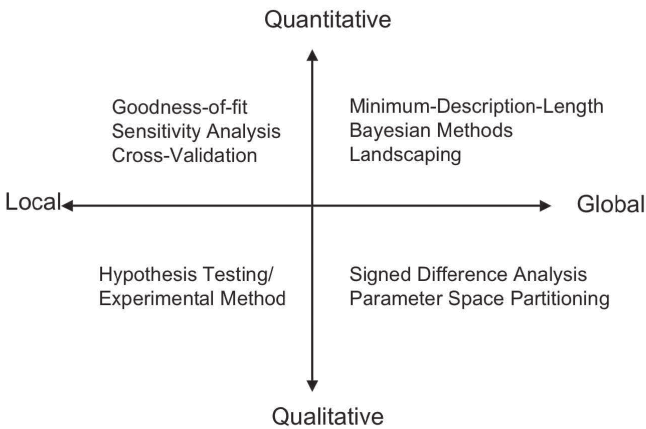
The dissimilarity of models, immediately apparent from the cartoons in Figure 6, can be traced to fundamental decisions about their architectures (E-Z Reader: stochastic automaton, SWIFT: stochastic dynamical system, SERIF: stochastic model, Glenmore: connectionist network/coupled difference equations) and to their implementation of temporal (when) and spatial (where) decisions. The fact that the models do a comparably good job in accounting for benchmark results with a comparable degree of model complexity as indexed by the number and constraints on free parameters, strongly suggests that the models are not sufficiently constrained by the benchmark results. Therefore, a more promising route than comparisons in terms of goodness of fit, is to increase the scope of results they are expected to simulate.

---

<sup>3</sup> Differences between Figure 5 and Figure 1 (which displays similar probabilities) are due to language differences between English and German. Data in Figure 1 were fit by the later version of SWIFT (Engbert et al., 2005).



**Figure 6:** Architectural blueprints of E-Z Reader (top left, from Reichle et al., 1998, Fig. 4), SWIFT (top right, from Engbert et al., 2005, Fig. 6), SERIF (bottom left, from McDonald et al., 2005, Fig. 1), and Glenmore (bottom right, from Reilly and Radach, 2006, Fig. 1) exhibit a high dissimilarity between these computational models of eye-movement control during reading.



**Figure 7:** Classification of methods for model analysis and comparison in a two-dimensional space (modified from Pitt et al., 2006), defined by the degree to which the method evaluates quantitative versus qualitative model performance (vertical axis) and whether the method focuses on local or global model behavior (horizontal axis).

Pitt, Kim, Navarro, and Myung (2006) suggest a classification of methods for model analysis and selection in a two-dimensional space, where the first dimension represents the range from local to global methods and the second dimension is a scale from qualitative methods to quantitative methods (Figure 7).

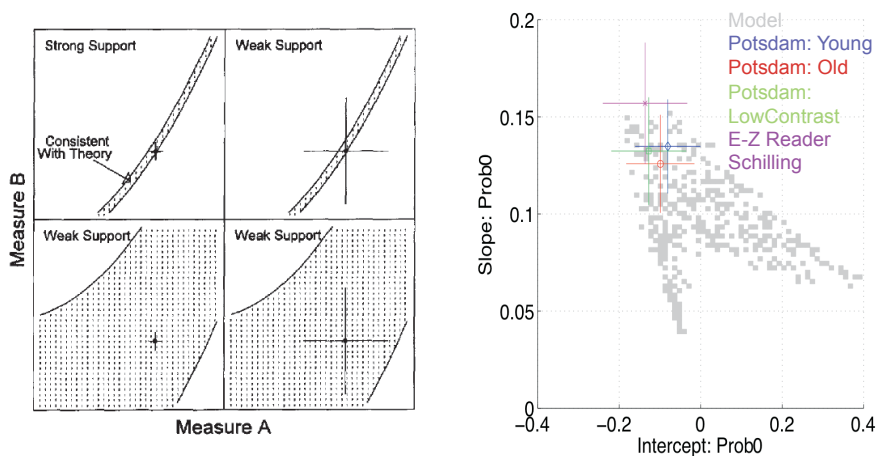
Psychological research almost exclusively applies local methods, but methods representing the global approach are needed to test model reliability and generalizability. The reason for this imbalance is that the applicability of quantitative global methods to the diverse range of models in psychology is currently “limited by their technical requirements” (Pitt et al., 2006). An important reason is that realistic models are computational rather than analytical, which creates problems for the implementation of methods of model analysis and comparison.

## 5 Meeting the Roberts and Pashler (2000) challenge

Global methods for data analysis, quantitative and qualitative, may become available in a convincing way for psychological models over the next years. Currently, we do not see a straightforward application for the models in our domain of research. In psychology, the problem is that we are usually satisfied with good model fits, rarely complemented by sensitivity analyses of parameters or cross-validation of results. Therefore, for now, we may remain within the local-quantitative quadrant of Figure 7. Roberts and Pashler’s (2000, 2002) starting point is that, although psychological models may live up to reasonable expectations about goodness of fit, goodness of fit does not discriminate convincingly between the models and is only a necessary condition for model evaluation. Consequently, they argued that there is currently no psychological theory in the sense in which we use this term in physics because psychological theories (or models) fall short on the following three criteria: strictness of model, reliability of data, and unexpected predictions. True or not, let us proceed from the assumption that several computational models of eye-movement control in reading, varying widely in theoretical assumptions and architectures, recover critical experimental benchmark results with the same number of free parameters, but cannot be distinguished in goodness of fit. In the following we review research and outline a research program that, at least in perspective, may allow us to meet this challenge for the SWIFT model.

## 5.1 Strictness of model and reliability of data

Figure 8 serves to illustrate the first two criteria: strictness of model and reliability of data. The left panel is taken from Roberts and Pashler (2000). Crosses with error bars represent two experimental measures or estimates A and B that can be derived from empirical data. Depending on the reliability of the experimental data, the error bars will be narrow (left column) or wide (right column). The grey points are results from model simulations; they are predictions of the models for different combinations of the model parameters. Thus, a strict model (top row) will generate a smaller set of predictions than a very flexible model (bottom row). In each panel, the data fall into the grey model zone. Thus, the models are always consistent with the data. By itself, however, the goodness-of-fit criterion does not say anything about model strictness and data reliability. The models in the bottom row are too flexible and the data in the right column are too variable. Strong support of the model is only present in the top left panel, where model strictness and reliability of data are in a reasonable relation with each other.



**Figure 8:** Left: Illustration of model strictness and reliability of data (from Roberts and Pashler, 2000, Fig. 1). Crosses represent experimental results; grey dots represent model simulations. Criteria for strict model and reliable data fulfilled for top-left panel. Right: Application to relation between intercept (x-axis) and slope (y-axis) for regression of skipping probability on log of word frequency. Crosses indicate different data sets; grey dots indicate results from SWIFT simulations.

As a concrete example from our research (see right panel of Figure 8), we regressed skipping probability on log of word frequency; from this we obtain two “mea-

tures”, the intercept and the slope (which is positive as skipping increases with word frequency). In a next step we carry out a large number of model simulations with parameters drawn randomly from reasonable ranges. From the data of each simulation we derive intercept and slope for the two measures. These values, again, are shown as grey dots and represent possible predictions by the SWIFT model. The crosses represent intercepts and slopes from different data sets, comprising young and old German readers, young German readers reading with strongly reduced screen contrast, and young English readers. These results show that the reliability of the estimates is quite comparable across data sets and more importantly that the between-language variation has a much stronger effect than the within-language age or within-language contrast manipulation. This suggests that language-comparative research in reading may hold much potential to explore the “legal” parameter settings of the model. Most importantly, however, we argue that these results suggest that the between-simulation variability of the SWIFT model is in a reasonable agreement with the experimental results. The next step is to expand the measurement space and, of course, to engage in systematic comparisons between models of claiming similar goodness of fit of benchmark results with similar number of model parameters.

## 5.2 Unexpected predictions

By far the toughest criterion to meet, the gold standard for a model is to generate predictions about behavior that is subsequently recovered from the data or experimentally established. Lakatos (1978, 6) put this succinctly:

The hallmark of empirical progress is not trivial verification. ... It is no success for Newtonian theory that stones, when dropped, fall towards the earth, no matter how often this is repeated. But so-called ‘refutations’ are not the hallmark of empirical failure, as Popper has preached, since all programmes grow in a permanent ocean of anomalies. What really counts are dramatic, unexpected stunning predictions: a few of them are enough to tilt the balance...

As an example for a stunning prediction of Newtonian physics, he mentions Halley’s exact prediction of space and time for the return of Halley’s comet 72 years later. Lakatos expresses the essence of the third problem: The *a priori* probability that the theory will fit the data is often ignored. At the end of this section we will give an example from our research, which was surprising to us. This example, however, is not representative of normal model development, given that none of the psychological theories we are aware of would seriously claim to be in the Newtonian league of scientific theories. Indeed, history of science regularly

uncovers the meandering between alternative conceptualizations and difficulties in choosing between them at the time of emergence of theories, which are now known only in a single canonical form.<sup>4</sup>

### 5.2.1 Case 1: Surprising results, incompatible with model predictions

Most frequently, development of psychological models is driven by new experimental results. Of course, in part this is simply due to the fact that there are many more experimental psychologists contributing new knowledge than there are modelers (and models) who can devote time to address the new results with their models. Indeed, we suspect that models will often not handle new experimental results adequately as they are implemented at the time of their publication. However, the results may not be incompatible with the theoretical principles guiding the model implementation (Rayner, 2009).

One example of such a result relates to fixation durations prior to skipped words. Experimentally, we observed that fixations before skipped words were shorter before short (or high-frequency) words (“skipping benefit”) and longer before long (or low-frequency) words (“skipping cost”; see Kliegl and Engbert, 2005, Kliegl, 2007, for details). The observation of skipping benefits is critical for models based on sequential attention shifts (SAS) like E-Z Reader (Reichle et al., 1998). In such a model, word skipping can only be produced by (i) cancellation of a saccade program to the next word  $n+1$  and (ii) the initiation of a new saccade program to word  $n+2$ . As a consequence, models of the SAS class always generate skipping costs, i.e., longer fixation duration before skipped words.

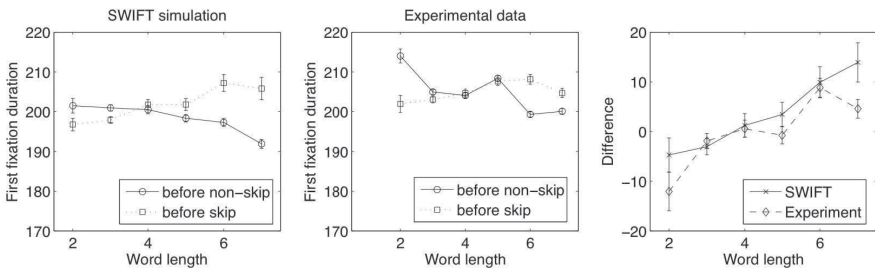
The SWIFT model (Engbert et al., 2005) also predicts skipping cost but this prediction was not tied as strongly to the model architecture as it is for the E-Z Reader model. The example is quite illustrative, because the skipping cost arises for a very different reason: the longer a fixation duration, the longer the preprocessing of the next word, and the higher the chances that the next word will be skipped. Thus, whereas in E-Z Reader long fixations prior to skipped words are a consequence of skipping, they are the cause of skipping in SWIFT. Nevertheless, neither model correctly recovered the skipping benefit associated with short words – counter to our own experimental data. Fortunately, in psychology, such reports of a falsification of a model do not necessarily preclude publication.

---

<sup>4</sup> For example, Damerow, Freudenthal, McLaughlin, and Renn (1992) describe and explain the problems and misunderstandings during the transition from early concepts of motion to the theory of motion in classical mechanics, using, among others, texts by Descartes and Galileo about the free fall of bodies and the composition of motions and forces.

### 5.2.2 Case 2: Surprising results, compatible with model after its modification

Box's (1979) "all models are wrong, some models are useful" is the guiding overarching principle. Falsifications are useful if they inspire model modification that encompasses new results in a principled way, rather than by some ad-hoc fix of the model. Indeed, such results frequently spur model modification to account for the results in such a way that ideally previous successful simulation results are preserved. Since the original publication of the failure to account for skipping cost, we have used this failure as one starting point for the further development of the model. In particular, we implemented the theoretical proposal of a dynamical modulation of the perceptual span, contingent on the foveal processing difficulty as described in section 2.2 (see also Figure 2, Eq. 3). We assumed that (i) the extension of the processing span to the left is constant, (ii) the processing span is symmetric for high foveal load, and (iii) the extension to the right increases with decreasing foveal load. Next, we fitted all model parameters of this variant of the SWIFT model using the same methods as reported by Engbert et al. (2005).



**Figure 9:** Skipping costs and benefits as a function of word length in experimental data and SWIFT simulations. The left panel shows average fixations durations in SWIFT simulations. The center panel shows the same plot for experimental data. The right panel shows the fixation durations before skipping subtracted by the fixation durations before non-skipping as a function of word length. The model simulations reproduce the experimental result that there are skipping costs for long words (word length > 5 letters) and skipping benefits for short words (< 4 letters).

Are there consequences of the dynamic processing span in SWIFT for the issue of word skipping discussed in the last section? Specifically, will this modification reveal skipping benefit prior to short words? Since parafoveal processing is very important to word skipping, it is plausible that the dynamical modulation of the perceptual span will lead to new results. Experimentally, we had observed that fixations before skipped words were shorter before short (or high-frequency)



words and longer before long (or low-frequency) words (Figure 9, center panel; see Kliegl and Engbert, 2005, for details). Interestingly, the SWIFT variant with dynamic foveal span can reproduce this highly specific data pattern accurately (Figure 9, left panel). The good agreement between experimental data and SWIFT simulations can be made visible, when differences in fixation durations (fixation durations before skipplings subtracted by fixation durations before non-skippling) are plotted (Figure 9, right panel).

These results from pilot simulations represent a major model improvement, because the current version of SWIFT always generated skipping costs (increased fixation durations before skipping) between 10 ms (word length 2) and 60 ms (word length 6). To our knowledge, the variant of the SWIFT model with dynamic foveal processing span investigated here is the only computational model that can reproduce the patterns of fixation durations before skipped words.

The model modification also “survived” two important tests. First, adding a new principle to an existing model might change the model’s performance on benchmark tests. Evaluations based on summary statistics for fixation durations and fixation probabilities, however, indicate that the dynamic processing span is as compatible with experimental data as a constant, asymmetric processing span. The overall goodness-of-fit of the model was not affected by the dynamic processing span. Second, the introduction of a dynamic processing span might have a strong impact on the effects of word properties of the last and next words (Kliegl et al., 2006), because variations of the extension of the processing span in general will change parafoveal processing. Interestingly, such changes in model performance were not observed from our pilot simulations. Thus, at this point the dynamical span served as defensible extension of the original SWIFT model. For a continuation of this story in the context of a further modification of the model we refer to Engbert and Kliegl (2011). We submit that this back and forth between experimental results and model development accounts for most of the research time in model development.

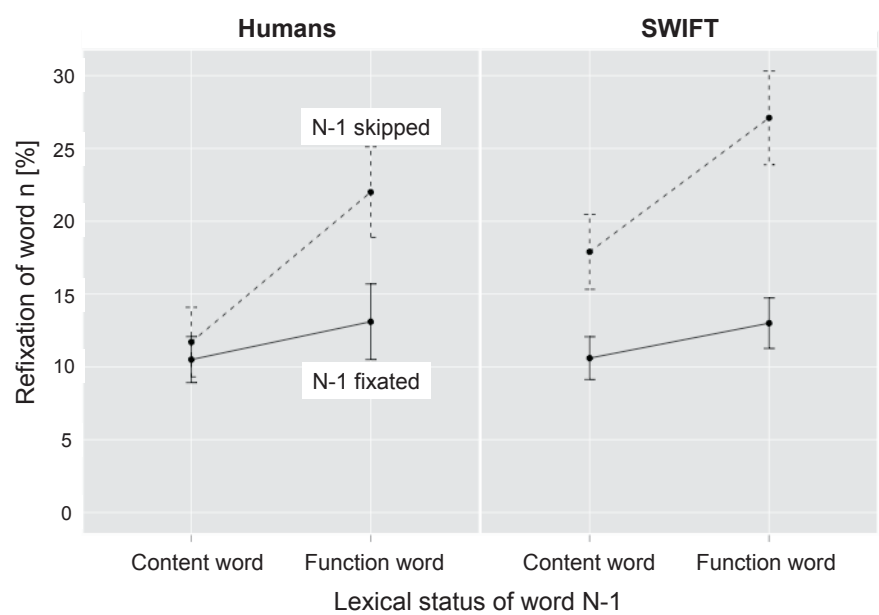
### 5.2.3 Case 3 (Lakatos Case): An unexpected prediction lurking in the data

We conclude with a “gold standard” example of an unexpected prediction derived from the SWIFT model and confirmed by the analyses of data collected many years ahead of the prediction (Risse et al., 2008): *Refixation probability should be larger after a skipped function word, but not after a skipped content word.*

The prediction is derived from the assumption that there is a competition between saccade targets. Suppose you are looking at a word and next to this word is a preposition. The preposition will likely be processed during this fixation,

because prepositions are among the most frequent and most predictable words. This implies that the next word drops out of the race to become selected as the next saccade target. At the same time the second word to the right will be slowly raising its activity level. So its chances to become elected increase compared to the situation with a competitor in position N+1. And most importantly, it will be fixated early in its activation profile, which increases the chances that it will be refixated.

The results are presented in Figure 10 and match the prediction very well. After a skipped function word the refixation probability is higher than after a skipped content word. There is some misfit, too: The overall refixation rate after skipping is overestimated.



**Figure 10:** Refixation rate after skipped and fixated content and function words. Left: Human data. Right: SWIFT simulation results.

Finally, the prediction lends itself to a comparison with the E-Z Reader model which also makes the general prediction that refixation rate will be higher after skipped words, but will generate the opposite prediction with respect to lexical status: Refixation rate should be *lower* after a skipped function word than a skipped content word. The reason for this prediction is that a skipped function

word provides longer preview of the following word than a skipped content word in this model. So there will be more need for processing after a skipped content word than after a skipped function word.

## 6 Perspectives and conclusion

### 6.1 Implications beyond reading

Experimental and mathematical psychology have developed detailed models of the interplay between cognitive subsystems (e.g., perception, attention, language, motor control). Dynamic models based on this approach can provide powerful theoretical blueprints for the behavioral and also for the neural organization of these cognitive processes,

- (1) if they are simulated on a computer with advanced techniques and studied qualitatively within the framework of nonlinear-science models,
- (2) if model parameters are estimated from high-resolution time series, and
- (3) if both experimental data and model simulations are evaluated by advanced methods for the analysis of complex multivariate time series.

As an example, we have described competing models that aspire to meet these criteria in the domain of reading research (Engbert et al., 2005; McDonald et al., 2005; Reichle et al., 1998; Reilly and Radach, 2006). There are hardly any more convenient measures than eye movements if one is interested in how behavior rapidly unfolds over time. Thus, eye movements represent an ideal model system in experimental psychology.

Most importantly, the neural circuits subserving the generation of eye movements are well understood (Sparks, 2002). Eye movements may well be the most direct behavioral signatures of neural firing, for they are directly related to spatio-temporal activation in the superior colliculus (SC). Indeed, the minimum of the oculomotor response time is about 60 ms after visual stimulus presentation, where the estimate is based on brainstem circuitry. More and more is currently learned about how higher-order structures (e.g., frontal eye fields, lateral intraparietal cortex, visual cortex) modulate brainstem nuclei when the oculomotor system is triggered by perceptual, attentional, and vestibular demands (e.g., Munoz and Everling, 2004). Further, because the loads on the extraocular muscles do not vary, reverse modeling can be used to reconstruct the eliciting innervation pattern. Most importantly, saccade and fixation parameters describing the eye movements across a visual scene or across a text embody behavioral dynamics

in experimental designs covering the broad spectrum of behavioral activity from simple perception via reading to postural control.

## 6.2 Model analysis and comparison

Starting with Huey (1908), research of eye movements in reading has been impressed with the range and stability of differences between individual readers as well as the magnitude of effects induced by differences in task demands. For example, individual differences in single-fixation durations among readers varying from 18 to 80 years of age account for more variance than 18 fixation-positions and psycholinguistic predictors (Kliegl et al., 2006). In agreement with the early research, preliminary analyses of data from bilingual readers of English and German varying widely in second-language proficiency suggest that individual differences will be even more pronounced in skipping and refixation probabilities. Thus, accounting for this variance in the SWIFT model would represent a major step in the further development of the model. We also note that there is no other computational model of reading or other cognitive processes that has been expanded in this direction.

Summary statistics relating to fixation durations and probabilities as a function of word length and word frequency can be reproduced remarkably well by at least four computational models of eye-movement control during reading. They all succeed with respect to the necessary condition of **goodness of fit** with a comparable number of free model parameters. Here we went beyond this necessary condition and offer some evidence that the SWIFT model may also live up to expectations of a strict set of criteria relating to model strictness, reliability of data, and unexpected predictions, as postulated by Roberts and Pashler (2000).

As a test of **model strictness** and **reliability of the data** we showed that the covariation of intercept and slope from the regression of word-skipping probability on log word frequency across simulations of the SWIFT model with random variation of model parameters within plausible ranges of parameter values agrees very well with variation observed between different reader groups varying in age, contrast of screen, and language.

The requirement of **unexpected model predictions** is illustrated in the form of three cases. First, in psychological research it is still more common to be surprised by new results. They may be compatible with model principles, but not recovered by a model in its current implementation. Second, some surprising results are bound to lead to constructive modifications of principles and implementations. Usually, these new “successes” are to be cumulative to earlier ones. Third, sometimes model predictions can be evaluated with respect to their

agreement with previously not known facts. Such predictions should have a low a priori probability in the scientific community; they must not be trivial.

Model modification usually changes model complexity and requires a consideration of its own. Neal (1996, 103–104) aptly summarized the issue of model complexity as follows:

Sometimes a simple model will outperform a more complex model ... Nevertheless, ... deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well.

In summary, we have contributed key findings to both experimental and computational aspects of eye-movement control during reading. We developed a computational model based on the assumption on distributed processing (SWIFT; Engbert, et al., 2002, 2005, Schad and Engbert, 2012). The model accounts for a large number of experimental observations, e.g., various measures of inspection probabilities and inspection durations, eye landing positions within words, delayed lexical access, parafoveal preprocessing. With respect to the scope of covered phenomena and transparency of its theoretical principles, arguably, SWIFT is currently the most advanced model of eye-movement control during reading. Of course, there are also aspects of reading behavior the model cannot and cannot be expected to get right at this point in time (see Engbert et al., 2005; Risse et al., 2008), but it certainly is a very useful tool guiding much of our research (Box, 1979).

## 6.3 Conclusion

We like to think about eye movements during reading as the “drosophila of psychological modeling” because they map onto a comparatively simple measurement space within which behavior of a surprisingly high level of complexity unfolds. It is a general critical requirement for modeling of cognitive processes to focus on a field of study with just the right level of complexity of behavior for the intended model. Eye movements during reading appear to meet this expectation in an ideal way.

## 7 Acknowledgement

This research was supported by Deutsche Forschungsgemeinschaft Research Group 868 “Computational Modeling of Behavioral, Cognitive, and Neural Dynamics” (Grant KL955/14) and European Science Foundation (ESF; Grant 05\_ECRP\_FP\_006, DFG KL955/7).

## References

- Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* 4. 91–99.
- Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In: Launer, R. L. & Wilkinson, G. N. (eds.), *Robustness in statistics*. New York: Academic Press.
- Damerow, P., Freudenthal, G., McLaughlin, P., & Renn, J. (1992). *Exploring the limits of preclassical mechanics*. New York: Springer.
- Engbert, R. (2007). *Reading with a dynamic processing span*. Presentation at 14<sup>th</sup> ECEM, Potsdam, Germany.
- Engbert, R. & Kliegl, R. (2011). Parallel graded attention models in reading. In: Liversedge, S. P., Gilchrist, I. & Everling, S. (eds.). *The Oxford Handbook of Eye-Movements*. New York, NY: Oxford University Press. 787–800.
- Engbert, R. & Kliegl, R. (2001). Mathematical models of eye movements in reading: A possible role for autonomous saccades. *Biological Cybernetics* 85. 77–87.
- Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research* 42. 621–636.
- Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review* 112. 777–813.
- Findlay, J. M. & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences* 22. 661–721.
- Henderson, J. M. & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16. 417–429.
- Huey, E. B. (1908). *The psychology and pedagogy of reading*. New York: Macmillan.
- Inhoff, A. W. & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics* 40. 431–439.
- Kennedy, A. (2003). The Dundee corpus [CD-ROM]. Dundee, Scotland: University of Dundee, Department of Psychology.
- Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, & Reichle. *Journal of Experimental Psychology: General* 136. 530–537.
- Kliegl, R. & Engbert, R. (2005). Fixation durations before word skipping in reading. *Psychonomic Bulletin & Review* 12. 132–138.

- Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology* 16. 262–284.
- Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General* 135. 12–35.
- Lakatos. I. (1978). The methodology of scientific research programmes. In: Worrall, J. & Currie, G. (eds.). *Philosophical Papers* 1. Cambridge: Cambridge University Press.
- McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., & Jacobs, A. M. (1989). Eye movement control during reading: II. Frequency of refixating a word. *Perception & Psychophysics* 46. 245–253.
- McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics* 17. 578–586.
- McDonald, S. A., Carpenter, R. H. S., & Shillcock, R. C. (2005). An anatomically constrained, stochastic model of eye movement control in reading. *Psychological Review* 112. 814–840.
- Munoz, D. P. & Everling, S. (2004). Look away: the anti-saccade task and the voluntary control of eye movement. *Nature Review Neuroscience* 5. 218–228.
- Neal, R. M. (1996) *Bayesian learning for neural networks*. New York: Springer.
- Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review* 113. 57–83.
- Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology* 62. 1457–1506.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124. 372–422.
- Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception* 8. 21–30.
- Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology* 7(1). 65–81.
- Rayner, K. & Pollatsek, A. (1987). Eye movements in reading: A tutorial review. In: Coltheart, M. (ed.). *Attention and Performance* 12. New York: Academic Press.
- Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T. J., & Reichle, E. D. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert (2006). *Journal of Experimental Psychology: General* 136. 520–529.
- Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research* 7. 4–22.
- Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review* 105. 125–157.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences* 26. 446–526.
- Reilly, R. & Radach, R. (2003). Foundations of an interactive activation model of eye movement control in reading. In: Hyönä, J., Radach, R., & Deubel, H. (eds.). *The mind's eye: Cognition and applied aspects of eye movement research*. Oxford: Elsevier. 429–455.
- Reilly, R. G., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research* 7. 34–55.
- Risse, S., Engbert, R., & Kliegl, R. (2008). Eye-movement control in reading: Experimental and corpus-analytic challenges for a computational model. In: Rayner, K., Shen, D., Bai, X.,

- & Yan, G. (eds.). *Cognitive and cultural influences on eye movements*. Tianjin: Tianjin People's Publishing House. 65–91.
- Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107. 358–367.
- Roberts, S. & Pashler, H. (2002). Reply to Roders and Rowe. *Psychological Review* 109. 605–607.
- Schad, D. J. & Engbert, R. (2012). The zoom lens of attention: Simulating shuffled versus normal text reading using the SWIFT model. *Visual Cognition*, 20(4–5). 391–421.
- Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition* 26. 1270–1281.
- Sparks, D. L. (2002). The brainstem control of saccadic eye movements. *Nature Review Neuroscience* 3. 952–964.
- Vitu, F., McConkie, G. W., Kerr, P., & O'Regan, J. K. (2001). Fixation location effects on fixation durations during reading: an inverted optimal viewing position effect. *Vision Research* 41(2526). 3513–3533.

**Prof. Dr. Reinhold Kliegl**

University of Potsdam  
Department of Psychology  
Karl-Liebknecht-Str. 24–25  
14476 Potsdam  
Germany  
kliegl@uni-potsdam.de

**Prof. Dr. Ralf Engbert**

University of Potsdam  
Department of Psychology  
Karl-Liebknecht-Str. 24–25  
14476 Potsdam  
Germany  
ralf.engbert@uni-potsdam.de



Martin Hoffmann

# Considering Criteria for Model Modification and Theory Change in Psychology

Commentary on Reinhold Kliegl and Ralf Engbert

## 1 Introduction

In contemporary philosophy of science, there are two main approaches to reconstructing the relation between models and theories. According to the so-called semantic view of theories, theories are just families of models. It was Patrick Suppes who declared that “a theory is a linguistic entity consisting of a set of sentences and models are non-linguistic entities in which the theory is satisfied” (Suppes 1960, p. 290); and this basic idea was elaborated by Sneed (1971), van Fraassen (1980), Stegmüller (1986), Suppe (1989), and others. Recently, this approach was criticised by some authors because it cannot account for the complex role that models play in scientific practice (Morrison and Morgan 1999; Suárez 1999). These authors propose a second, alternative approach to the relation between models and theories which regards models as almost independent from theories. Models are “autonomous agents” (Morrison 1999) that mediate between the level of general theories and concrete empirical data, and fulfil a variety of functions: models are useful instruments to represent reality, but also to test, explore, and elaborate theories.

Kliegl and Engbert outline a picture of modeling that fits the second approach of the model/theory-relation. In their paper they discuss many interesting and innovative ideas how to enhance the methods and how to expand the criteria for model evaluation in psychology. The main aim of this endeavour is to gain a powerful tool which will enable us to identify the best model amongst the many alternatives suggested in the present psychological discussion.

Kliegl and Engbert’s account is inspired by a radical critique of the current practices of model evaluation in psychology, which have been formulated by Roberts and Pashler (2000, 2002). The main point of Roberts and Pashler’s challenge is the following: Many theories in psychology are primarily tested and confirmed by their ability to fit the data. The methodological discussion focuses on developing methods to determine the goodness of fit. But, say Roberts and Pashler, goodness of fit alone is not sufficient to determine the empirical validity and the explanatory power of a model. They propose three additional criteria, namely model strictness, reliability of data and unexpected predictions. In their

paper, Kliegl and Engbert apply these criteria in a subtle way to different versions of their SWIFT model. They show that model strictness and reliability of data can be measured at least in principle, and that they can help to compare the adequacy of SWIFT with competing accounts. They identify the third criterion of Roberts and Pashler's as the main difficulty: unexpected, but correct model predictions. On the one hand, Kliegl and Engbert qualify this criterion as the "gold standard for a model" (section 5.2). On the other hand, they conclude that the SWIFT model meets this criterion only in some cases. It is more common that, in the beginning, there are surprising experimental findings. Then these findings are compared with model predictions, and one has to test whether the model can fit the data – which often requires a suitable modification. The central question is this: is such a modification of the model legitimate? Or should its failure to generate a correct prediction be regarded as a strong reason to abandon the model? I can only see two possible solutions: either the third criterion by Roberts and Pashler is unreasonably strict, or it is methodologically problematic to adhere to a model which generates wrong predictions.

My aim in this commentary is to discuss this tension on the basis of concrete examples from the SWIFT model. But before I will do that, it is necessary to say something about the epistemological reasons for the third criterion.

## 2 Why are unexpected, but correct model predictions important at all?

Roberts and Pashler take the idea for their third criterion from Imre Lakatos' philosophy of science. Kliegl and Engbert quote Lakatos' claim that not all predictions, but only the "dramatic, unexpected, stunning predictions" can corroborate the theory in question (Lakatos 1978a, p. 6). But why should it be important that an empirical result is novel, unexpected, or even stunning? These seem to be merely psychological categories, and it is not at all clear why such emotive responses should have any impact on the corroboration of theories and models. In order to see why these features are of any epistemological importance, it is necessary to make some remarks about Lakatos' philosophy of science.

Lakatos has formulated a re-statement of Popper's well-known idea of falsification as a criterion for the validity of empirical theories. Popper himself thought that success in science is primarily determined by strict tests of theories. A theory is falsified if a conflict with empirical data is indicated. In Lakatos' view, this version of falsificationism is naïve, so he puts his own version of sophisticated falsificationalism forward. Lakatos' main criticism is that Popper has construed

the central concepts of theory acceptance and falsification by employing two-place relations: observational data on the one hand, and the theory on the other hand. (By the way, in this respect there is a resemblance between Popper's ideas and the criterion of goodness of fit – which means fitting just one theory or model to a data structure). Lakatos thinks that this reconstruction is inadequate because scientific progress is only possible if a falsified theory is replaced by a promising successor. For this reason, in Lakatos' view, the competition between rival theories is neither an accidental property of science nor an indicator of a crisis; it is rather an essential element of fruitful theory development. Popper's two-place relation has – in Lakatos' conception of sophisticated falsificationalism – been replaced by a three-place relation between observational data and at least two rival theories  $T$  and  $T'$ .

A theory  $T$  is falsified if and only if a different theory  $T'$  is proposed which exhibits the following characteristics:

- (1)  $T'$  has excess empirical content over  $T$  ...;
- (2)  $T'$  explains the previous success of  $T$  ...
- (3) some of the excess content of  $T'$  is corroborated. (Lakatos 1978b, p. 32)

The requirement that  $T'$  has to be in accordance with novel, unexpected facts is essential for condition (1): "Excess empirical content" does not only mean that the area of application is broader, or the class of predictions bigger, than that of  $T$ . What is important here is that  $T'$  predicts facts that are improbable or forbidden according to  $T$ . This is precisely what constitutes the distinction between well-known and novel, unexpected facts relevant in the present context. Known and expected facts are facts that are in accordance with both theories  $T$  and  $T'$ . These facts are uninteresting for testing the new theory, because on their basis alone no decision between  $T$  and  $T'$  is possible. In contrast to that, reference to novel and unexpected facts (like the return of Halley's Comet) is decisive, because their confirmation allows for a justified choice between  $T$  and  $T'$ . The confirmation of correct but unexpected predictions is the decisive criterion to identify the theory with excess empirical content. But if one changes the content of the theory  $T'$  in the light of new data, the decision between  $T$  and  $T'$  becomes arbitrary, since it is possible to immunize every theory against conflicting data by making ad hoc assumptions.

For this reason, modifications in reaction to new data are problematic. If one allows for modifying  $T'$  in a way to accommodate the new data, Lakatos' theory of justified theory choice no longer works. Kliegl and Engbert point out that model modifications in fact do play an important role in current model evaluation. So the question arises whether one can give a rational reconstruction of this methodological procedure. Let us have a closer look at Kliegl and Engbert's examples.

### 3 Examples

Kliegl and Engbert present an example that fulfils the third criterion: concerning refixation probabilities after skipped content and function words, the SWIFT model generates unexpected but correct predictions (section 5.2.3). But they themselves admit that this example is “not representative of normal model development” (section 5.2). Usually, a process of model modification is initiated in the light of new experimental data.

Kliegl and Engbert present two examples to illustrate this process (sections 5.2.1 and 5.2.2). They focus on the fixation durations prior to skipped words: in the majority of cases word skipping generates a *skipping cost*, that means an increase of the fixation duration before skipped words compared to the fixation duration before fixated words. But experimental findings surprisingly show that there are also *skipping benefits* (that is *decreased* fixation durations before skipped words), which occur when the words are short or occur with high frequency.

These skipping benefits conflict with the predictions of the initial SWIFT model, presented by Engbert et al. (2005). But in contrast to alternative models for eye movements during reading (like the E-Z-Reader model), the SWIFT model’s predictions can be altered by varying the value of a free model parameter. The SWIFT model can account for a dynamical modulation of the perceptual span depending on word length. This modification allows for fitting the model to the experimental data. The modified SWIFT model is in accordance with the experimental data and predicts skipping benefits when the words in question are short. This methodological move deserves careful interpretation. Let me analyse in more detail which function the model modification might have in this particular context.

First of all, it has to be said that the better fit to the data generated by the modification contributes nothing to the corroboration of the model. Its initial predictions went wrong and it is re-stated given the conflicting data. So it does not exceed the empirical content of the old model in this respect. But nevertheless there is an important difference to a mere ad hoc hypothesis: the model assumptions are not just relaxed, but changed. This means that the old model predictions are partly replaced by new ones. Even if this does not corroborate the model, it may turn out theoretically fruitful: the new predictions can at least potentially be confirmed by new, unexpected empirical data. In fact, this is the case in Kliegl and Engbert’s second example. They report that the invention of a dynamic perceptual span in the SWIFT model does not affect the goodness of fit in one other important respect (section 5.2.2). In a sense, new and successful predictions can outweigh the model relaxation caused by the model modification.

But perhaps the focus on the predictive power of models is too narrow. There may be other aspects that have to be taken into account to describe a model's methodological role. In order to explain this, I would like to draw attention to the distinction introduced at the beginning: the distinction between models and more general and unified theories. It is important to notice that Lakatos only considers research programmes on the level of *theories*. On this general level, the methodological aim is to replace one theory by a progressive successor. But it is questionable whether the relation between a model and its successor has to be reconstructed in the same way. Following Morrison (1999), models are “autonomous agents” and for this reason it is an oversimplification to identify series of models with series of succeeding theories. According to this view, models are specific, flexible instruments that can be characterized independently of the underlying theories and that can be used for a variety of methodological purposes. But this type of “autonomy” of models does not imply that the theories themselves are irrelevant. On the contrary, the development of models is no end in itself. Models are rather designed as instruments or tools for developing more unified theories. They simply do not only serve to corroborate the theory, but also to explore the theory, develop new predictions, apply the theory to special and new areas, etc. So in order to integrate models into Lakatos' picture of the development of research paradigms, it is important to clarify the relations that obtain between models and general theories.

Therefore, in the present context the following question becomes central: what is the theory in question that should be refined by the modifications of the model? Kliegl and Engbert do not say much about the relation between the SWIFT model and a corresponding theory. But I think that the conflict between the two following theories lies at the heart of their project: their aim is to confirm the theory of parallel or distributed processing and to argue against a theory of serial or sequential processing. Serial processing is characterized by two assumptions: (i) attention is focused on just one word at a time and (ii) attention shifts mandatorily from one word to the next. Distributed processing is characterised by loosening both assumptions: (i') attention is a process allocated parallel on different words, which is called an activation field, (ii') attention shifts are explained by different patterns. These general theories are not restricted to eye movement control, but claim to give unified accounts of motor behaviour in general. Only one of these theories corresponds to the SWIFT models: the theory of distributed processing. For this reason the implementation of a dynamical perceptual span is possible in the SWIFT model, but it is not in line with models like the E-Z-Reader, formulated on the basis of the serial processing theory. Against this background, we can interpret the central function of the model modification in question –

while fitting the data – to generate new predictions for future confirmations of the theory.

Because of the complex relation between the SWIFT model and the distributed processing theory, it is difficult or impossible to define strict normative standards for particular model changes at the present time. But perhaps one can formulate two modest requirements for the relation between theory and models instead. First, it is of crucial importance that every model modification remains in accordance with the core assumptions of the theory. Second, every model modification should be defined such that the predictions of the new model still contradict the predictions of the competing theory. Only if these conditions are fulfilled, the modified model remains a useful instrument for developing the theory in question. In this particular case the model predictions are indeed in conflict with the predictions of the competing theory of serial processing, because the skipping benefits are extremely difficult to explain assuming serial processing. Serial processing theory explains word skipping as a termination of the saccade program if the next word is recognized. In this case the saccade program is cancelled and restarted to fixate the next but one word. But this process can only lead to skipping costs, not to skipping benefits. So the results of the modified SWIFT model, which predicts skipping benefits, contradict the theory of serial processing, but are perfectly in line with the theory of distributed processing.

## 4 Conclusion

To sum up, I am in favour of the following position concerning the plausibility of Roberts and Pashler's third criterion: prediction of unexpected, but correct empirical results. Applied to models, it is obviously too strong in its unrestricted formulation. It would be even irrational to require the fulfilment of this restrictive criterion as a necessary condition for accepting a model as part of a progressive research programme. However, this does not speak against Lakatos' ideas about strict theory testing. One has to consider that Lakatos applies this criterion only to the level of general *theories*, not to models. For this reason it is perfectly in line with Lakatos' account to maintain the third criterion as a plausible methodological rule for theories, and to abandon it for models.

In Lakatos' account it remains underdetermined how to react to conflicting data on the level of models. Kliegl and Engbert present many appealing and original ideas concerning model modifications in the particular case of the SWIFT model, but it remains difficult to derive more general rules from these. For now, it is merely possible to formulate one modest methodological restriction: model

modifications, which *only* relax the empirical content of a model, are problematic, because of their ad hoc character. The class of empirical data, which are in accordance with the model after its modification, should be outweighed by a class of data which were compatible with the previous model, and are not in line with its predictions after modification. This restriction is important to prevent mere ad hoc modifications which might render the model compatible with *every* experimental result.

So, primarily, this commentary is not intended to be a criticism of Kliegl and Engbert's application of the criteria by Roberts and Pashler, but rather a criticism of the application of methodological criteria for strict theory testing in the evaluation of modern, complex models. If one adopts the view of models as autonomous agents – suggested by Morrison, Morgan and others –, it remains an important task for future research in philosophy of science to clarify the complex relations between theories and models in more detail and to define precise methodological rules for how to modify models in the light of new data.

## References

- Engbert, Ralf, Nuthmann, Antje, Richter, Eike M., & Kliegl, Reinhold (2005). SWIFT: A Dynamical Model of Saccade Generation During Reading. *Psychological Review* 112. 777–813.
- Fraassen, Bas C. van (1980). *The Scientific Image*. Oxford: Oxford UP.
- Lakatos, Imre (1978a). Introduction: Science and Pseudoscience. In: Currie, John Worrall Gregory (ed.). *The Methodology of Scientific Research Programmes. Philosophical Papers* 1. Cambridge: Cambridge UP. 1–7.
- Lakatos, Imre (1978b). Falsification and the Methodology of Scientific Research Programmes. In: Currie, John Worrall Gregory (ed.). *The Methodology of Scientific Research Programmes. Philosophical Papers* 1. Cambridge: Cambridge University Press. 8–101.
- Morrison, Margaret (1999). Models as Autonomous Agents. In: Morgan, Mary S. & Morrison, Margaret (ed.). *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press. 38–65.
- Morrison, Margaret & Morgan, Mary S. (1999). Models as Mediating Instruments. In: Morgan, Mary S. & Morrison, Margaret (ed.). *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press. 10–37.
- Roberts, Seth & Pashler, Harold (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107. 358–367.
- Roberts, Seth & Pashler, Harold (2002). Reply to Roders and Rowe. *Psychological Review* 109. 605–607.
- Sneed, Joseph D. (1971). *The Logical Structure of Mathematical Physics*. Dordrecht: Reidel.
- Stegmüller, Wolfgang (1986). Theorie und Erfahrung: Dritter Teilband. Die Entwicklung des neuen Strukturalismus seit 1973. *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie* II/3. Berlin: Springer.

- Suárez, Mauricio (1999). Theories, Models, and Representations. In: Magnani, Lorenzo, Nersessian, Nancy J. & Thagard, Paul (ed.). *Model-Based Reasoning in Scientific Discovery*. New York; Boston; Dordrecht: Kluwer. 75–83.
- Suppes, Patrick (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese* 12. 287–301.
- Suppe, Frederick (1989). *The Semantic View of Theories and Scientific Realism*. Urbana; Chicago: University of Illinois Press.

**Dr. Martin Hoffmann**

University of Hamburg  
Department of Philosophy  
Von-Melle-Park 6  
20146 Hamburg  
Germany  
martin.hoffmann@uni-hamburg.de



Wolfgang Marquardt

# Identification of Kinetic Models by Incremental Refinement<sup>1</sup>

## 1 Introduction

Chemical (or process) engineering is an engineering science that focuses on the foundations of any kind of transformation of matter in order to change its molecular or morphological constitution. The primary subject of modeling is a (part of a) complete production process, which converts raw materials in desired chemical products. Any such (sub-)process comprises a set of connected pieces of equipment (or process units), which are typically linked by material, energy and information flows. The overall behaviour of the plant is governed by the behaviour of its constituents and their nontrivial interactions. The process can be considered as a system of systems (Marquardt, 1995): this process system forms a collection of subsystems, i.e., the pieces of equipment, which are connected by different types of flows forming a complex network. Every piece of equipment is structured itself; hence, its decomposition into interconnected subsystems is facilitated. Each of these subsystems is governed by typically different types of kinetic phenomena, such as (bio-)chemical reactions or intra- and interphase mass, energy and momentum transport. The resulting spatio-temporal behaviour is often very complex and yet not well-understood. This is particularly true if multiple, reactive phases (gas, liquid or solid) are involved.

Mathematical models are in the core “of methodologies for chemical engineering decisions (which) should be responsible for indicating how to plan, how to design, how to operate, and how to control any kind of unit operation (e.g., process unit), chemical and other production process and the chemical industries themselves” (Takamatsu, 1983). Given the multitude of model-based engineering tasks, any modeling effort has to fulfil specific needs asking for different levels of detail and predictive capabilities of the resulting mathematical model. While modeling in the sciences aims at an understanding and explanation of observed system behaviour in the first place, modeling in engineering is an integrated part of model-based problem solving strategies aiming at planning, designing, operating or controlling an artificial (process) system. There is not only a diversity of engineering tasks but also an enormous diversity of structures and phenomena

---

<sup>1</sup> This paper is based on previous reviews on the subject (Marquardt, 2005; Bardow and Marquardt, 2009).

governing (process) system behaviour. Engineering problem solving is faced with such multiple dimensions of diversity. A kind of “model factory” has to be established in industrial modeling processes in order to reduce the cost of developing models of high quality, which can be maintained across the plant lifecycle (Marquardt et al., 2000).

Models of process systems are multi-scale in nature. They span from the molecular level with short length- and time-scales to the global supply chain involving many productions plants, warehouses and transportation systems. The major building block of a model representing some part of a process system (sometimes also called a balance envelope) is the differential balance equation, which is formulated for a selected set of extensive quantities (Bird et al., 2002). The balances constitute of hold-up, of transport and source terms which reflect the molecular behaviour of matter on the continuum scale. Averaging is often applied to coarse-grain the resolution of the model in time and space for complexity reduction (Slattery, J. C., 1999). The bridging from the molecular to the continuum scale by some kind of coarse-graining results unavoidably in so-called closure problems. Roughly speaking, a closure problem arises, because the application of linear averaging operators to a nonlinear expression in a balance equation cannot be evaluated analytically to relate the average of such an expression to the averaged state variables (such as velocity, temperature, concentrations). The closure condition refers to some constitutive (in some cases even differential equation) model which relates the average of a nonlinear expression to the averaged state variables. A well-known closure problem refers to the determination of the Reynolds stress tensor which results from averaging the Navier-Stokes equations with respect to time (Pope, 2000). Even if such closure conditions are derived from theoretical considerations using some kind of scale-bridging approach, they typically require the identification of empirical parameters in the sub-model structures or in extreme cases even the model structure (i.e., the mathematical expressions relating dependent and independent variables) itself. In particular, the so-called  $k$ - $\epsilon$ -model for the Reynolds stress tensor comprises a number of parameters which have to be determined from experiments (Bardow et al., 2008).

Since such model identification is a complex systems problem, a goal-oriented work process has to be established which systematically links high resolution measurement techniques, mathematical modeling, real (laboratory) or virtual (simulation) experiments (typically on a finer scale) with the formulation and solution of so-called inverse problems (Kirsch, 1996). These inverse problems come in different flavours: they may be used to design the most informative experiment by fixing the experimental conditions in a given experimental set-up appropriately (Walter, Pronzato, 1990; Pukelsheim, 2006), to estimate paramete-

ters (Bard, 1974; Schittkowski, 2002) in a given model structure or to discriminate among model structure candidates based on experimental evidence (Verheijen, 2003). Typically, the model identification task cannot be successfully tackled in one go. Rather, some kind of iterative refinement strategy is intuitively followed by the modeller to exploit the knowledge gained during the model development procedure. Probably the most important decision to be made is the level of detail to be included in the target model to result in a desired model resolution.

To this end, this contribution summarizes recent progress towards a systematic work process (Bardow and Marquardt, 2004; Marquardt, 2005) to derive valid mathematical models for kinetically controlled reaction and transport problems, which govern the behaviour of (bio-)chemical process systems. This work process is called *model-based experimental analysis* (or MEXA for short) and aims at *useful models at minimal engineering effort*. While mathematical models of kinetic phenomena can in principle be developed using standard statistical techniques including nonlinear regression (Bard, 1974) and multi-model inference (Burnham, Anderson, 2002), this direct approach typically results in strongly nonlinear and large-scale mathematical programming problems (Schittkowski, 2002; Biegler, 2010), which may not only be computationally prohibitive, but also result in models which are not capturing the underlying physico-chemical mechanisms appropriately. In contrast, *incremental model identification* (or IMI for short), which is an integral part of the MEXA methodology, constitutes a physically motivated divide-and-conquer strategy to kinetic model identification.

This paper is structured as follows: Section 2 presents a general overview on the MEXA methodology. Two identification strategies, simultaneous and incremental model identification are introduced in Section 3. Sections 4, 5 and 6 sketch the application of the MEXA and IMI methodologies exemplarily to three challenging and relevant process modeling problems including (bio-)chemical reaction kinetics in single- and multi-phase systems, multi-component diffusion in liquids and energy transport in wavy falling film flows. The final Section 6 provides a summarizing discussion.

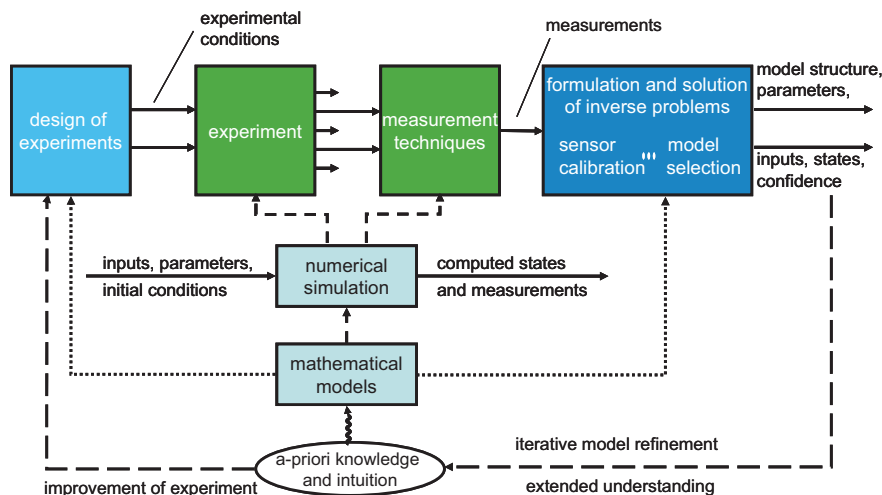
## 2 Model-based experimental analysis

An overview on the MEXA methodology is presented in Fig. 1. The typical workflow involves the following steps:

1. An initial experiment (comprising the experimental apparatus and appropriate measurement devices) is built to observe a kinetic phenomenon of interest.

2. Experimental evidence and the available a-priori knowledge are used to build a first structured mathematical model of the experiment. Unknown parameters are initialized with plausible values.
3. Virtual experiments are carried out by means of simulation studies using this first model. Even if the simulation results were only qualitatively correct, they will provide insight into the behaviour of the experiment prior to actual laboratory work, which could result in a revision of the design of the initial experiment and its operation.
4. First experiments are performed. They should be guided by statistical design of experiments (Mason et al., 2003) to explore a telling set of experimental conditions. These experiments will provide some qualitative insight into the behaviour of the experiment and the governing kinetic phenomena.
5. The measurement data recorded in the initial experiments can be used to formulate a parameter estimation problem, which is a special kind of inverse problem, for the conjectured model structure. Not all parameters may be identifiable. Therefore, parameter estimation should be preceded by identifiability analysis (Vaida et al., 1989, Walter, Pronzato, 1997) to assess which (combinations of) parameters can be estimated uniquely from the available measurements.
6. The model of the experiment is now used to find experimental conditions by means of optimal design of experiments (Walter, Pronzato, 1990; Pukelsheim, 2006) resulting in most informative data for the intended purpose of the experimental investigation. Such a revision of the experiment targets, in the first place, the operating conditions of the experiment and the type and accuracy of the measurements taken. However, also the experimental set-up is subject to possible change.
7. The designed experiment is performed and the observations are recorded. One or more inverse problems are formulated and solved to calibrate sensors, to estimate unknown inputs, states or parameters of the model or to select and discriminate an appropriate model structure.
8. Most often, the resulting model does not reflect the kinetic phenomenon of interest with sufficient detail and accuracy. In particular, the selected model structure may not properly match reality sufficiently well, or, the model may be too detailed to allow for its identification. The accumulated understanding, however, allows for an iterative improvement of the model, either by model simplification to improve identifiability (Quaiser et al., 2011) or by model structure refinement to better capture reality (Verheijen, 2003).
9. The sequence of steps 6, 7 and 8 is repeated until a model is obtained, which is fully consistent with all the measurements available. The investigations

should ideally only be terminated if the model can not be falsified by any conceivable experiment (Popper, 1959).



**Figure 1:** The MEXA methodology – an overview (Marquardt, 2005).

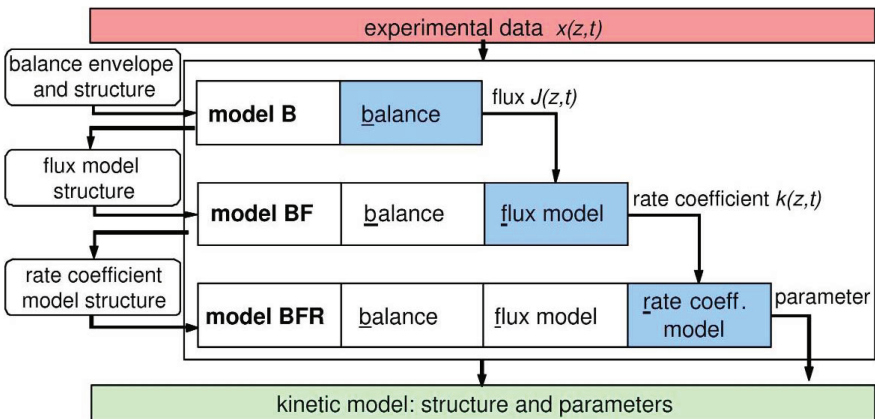
Research on systematic work processes for mathematical model development, which combine experiments, data analysis, modeling and model identification, dates back at least to the 1970s (Kittrell, 1970). However, the availability of current, more advanced experimental and theoretical techniques offers new opportunities to develop more comprehensive modeling strategies, which are widely applicable to a variety of modeling problems. For example, a modeling process with a focus on optimal design of experiments has been reported by Asprey and Macchietto (2000). The collaborative research centre CRC 540, “Model-based Experimental Analysis of Fluid Multi-Phase Reaction Systems” (cf. <http://www.sfb540.rwth-aachen.de/>), which was funded by the German Research Foundation (DFG), addressed the development of advanced modeling work processes comprehensively from 1999 to 2009. The research covered the development of novel high resolution measurement techniques, efficient numerical methods for the solution of direct and inverse reaction and transport problems and the development of a novel, experimentally driven modeling strategy which relies on iterative model identification.

### 3 Two alternative model identification strategies

This iterative model identification strategy is introduced in the following section and compared to the established simultaneous model identification (SMI) strategy.

#### 3.1 Incremental model identification

Incremental model identification (IMI) relies on an incremental refinement of the model structure which is motivated by systematic model development (Fig. 2) as suggested by Marquardt (1995). The major model development steps and their relation to incremental model identification are outlined in the following.



**Figure 2:** Incremental modeling and identification (Marquardt, 1995, 2005)

##### 3.1.1 Model B

In model development, balance envelopes and their interactions are determined first to represent a certain part of the system of interest. The spatio-temporal resolution of the model is decided in each balance envelope, e.g., the model may or may not describe the evolution of the behaviour over time  $t$  and it may or may not resolve the spatial resolution in up to three space dimensions  $z$ . Those extensive quantities  $y(z,t)$  are selected for which a balance equation is to be formulated. In case of spatio-temporally resolved models, the balance reads as

$$\begin{aligned}\frac{\partial y}{\partial t} &= -\nabla_z \cdot J_{f,y} + J_{s,y}, \quad y(z, t_0) = y_0(z), \quad z \in \Omega, \quad t > t_0, \\ \nabla_z y \Big|_{\Gamma} &= J_{b,y}, \quad z \in \Gamma,\end{aligned}\tag{1}$$

with  $y$  a selected extensive quantity (such as mass, mass of a certain chemical species, energy, etc.) propagated according to the transport term  $J_{f,y}(z, t)$  and generated (or consumed) according to the source term  $J_{s,y}(z, t)$  at any point in the interior of the balance envelope  $\Omega \subset R^n, n = 1, 2, 3$ . The symbol  $J_{b,y}(z_b, t)$  refers to transport across the boundary  $G$  of the balance envelope. Any extensive quantity  $y(\cdot)$  is related to a set of measured (typically intensive) quantities  $x(\cdot)$  by some constitutive relation

$$y(\cdot) = h(x(\cdot), \cdot).\tag{2}$$

If no spatial resolution of the state variables is desired, the balance for some  $y(t)$  is written as

$$\frac{dy}{dt} = j_{f,y} + j_{s,y}, \quad y(t_0) = y_0\tag{3}$$

The symbols  $j_{f,y}(t)$  and  $j_{s,y}(t)$  refer to transport of  $y(t)$  and its generation (or consumption) within the balance envelope, respectively.

Note that no constitutive equations are considered yet to specify either of the fluxes  $J_{f,y}$ ,  $J_{s,y}$ ,  $J_{b,y}$ ,  $j_{f,y}$  or  $j_{s,y}$  as a function of the intensive thermodynamic state variables  $x(\cdot)$ . While these constitutive equations are selected on the following decision level, unknown fluxes  $J_{\dots}$  or  $j_{\dots}$  are estimated directly from the balance equation in incremental model identification. For this purpose, measurements of  $x(\cdot)$  with sufficient resolution in time  $t$  and/or space  $z$  are assumed. An unknown flux,  $J_{\dots}$  or  $j_{\dots}$ , can then be estimated from one of the balance equations as a function of time and/or space coordinates without specifying a constitutive equation.

### 3.1.2 Model BF

In model development, constitutive equations are specified for each flux term in the balances on the next decision level. In particular,

$$\begin{aligned}
J_{f,y}(z,t) &= G_{f,y}(x, \nabla_z x, \dots, k_{f,y}), \\
J_{s,y}(z,t) &= G_{s,y}(x, \dots, k_{s,y}), \\
J_{b,y}(z,t) &= G_{b,y}(x, \nabla_z x, \dots, k_{b,y})
\end{aligned} \tag{4}$$

for spatially distributed and

$$\begin{aligned}
j_{f,y}(t) &= g_{f,y}(x, \dots, k'_{f,y}), \\
j_{s,y}(z,t) &= g_{s,y}(x, \dots, k'_{s,y})
\end{aligned} \tag{5}$$

for spatially lumped balance envelopes. The symbols  $k_{\dots}$  and  $k'_{\dots}$  refer to some rate coefficient functions which depend on time and space (in case of spatially distributed balance envelopes) or on time alone (in case of spatially lumped balance envelopes), respectively. These constitutive equations could, e.g., correlate interfacial fluxes or reaction rates with state variables.

Similarly, in incremental model identification, flux model candidates, as in Eqs. (4) or (5), are selected or generated on decision level *BF* to relate the flux to rate coefficients, to measured states, and possibly to their derivatives. The estimates of the fluxes  $J_{\dots}(z,t)$  or  $j_{\dots}(t)$  obtained on level *B* are now interpreted as *inferential* measurements. Together with the *real* measurements  $x(z,t)$  or  $x(t)$ , one of these flux estimates can then be used to determine one of the rate coefficients  $k_{\dots}$  or  $k'_{\dots}$  as a function of time and space from the corresponding equation in (4) or (5), respectively. Often, the flux model can be analytically solved for the rate coefficient function  $k_{\dots}(z,t)$  or  $k'_{\dots}(t)$ . These rate coefficient functions, for example, refer to heat or mass transfer or reaction rate coefficients.

### 3.1.3 Model BFR

In many cases, the rate coefficients  $k_{\dots}(z,t)$  or  $k'_{\dots}(t)$  introduced in the correlations on level *BF* depend on the states  $x(\cdot)$  themselves. Therefore, a constitutive model

$$\begin{aligned}
k_{\dots} &= r_{\dots}(x, \nabla_z x, \dots, \theta), \\
k'_{\dots} &= r_{\dots}(x, \dots, \theta),
\end{aligned} \tag{6}$$

relating the rate coefficients to the states, has to be selected on yet another decision level named *BFR* (cf. Fig. 2).

Mirroring this last model development step in incremental model identification, a model for the rate coefficients has to be identified. The model candidates, cf. Eq. (6), are assumed to only depend on the measured states, their spatial gra-



dients and on *constant* parameters  $q \in R^p$ . If only a single candidate structure is considered, the parameters  $q$  can be computed from the estimated functions  $k_{\dots}(z,t)$  or  $k'_{\dots}(t)$  and the measured states  $x$  by solving a (typically nonlinear) algebraic regression problem. In general, however, a model discrimination problem has to be solved, where the most suitable model structure is determined from a set of candidates.

The cascaded decision making process in model development and model identification has been discussed for three levels which commonly occur in practice. However, model refinement can continue as long as the sub-models of the last model refinement step not only involve constant  $q$  as in Eqs. (4)–(6), but rather coefficient functions, which depend on state variables. While this is the decision of the modeller, it should be backed by experimental data and information deduced during incremental identification such as the confidence in the selected model structure and its parameters (Verheijen, 2003).

This structured modeling approach renders all the individual decisions completely transparent, i.e., the modeller is in full control of the model refinement process. The most important decision relates to the choice of the model structures for the flux expressions and the rate coefficient functions in Eqs. (4)–(6). These continuum models do not necessarily have to be based on molecular principles. Rather, any mathematical correlation can be selected to fix the dependency of a flux or a rate coefficient as a function of intensive quantities. A formal, semi-empirical but physically founded kinetic model may be chosen which at least to some extent reflects the molecular level phenomena. Examples include mass action kinetics in reaction modeling (Higham, 2008), Maxwell-Stefan theory of multi-component diffusion (Taylor, Krishna, 1993) or established activity coefficient models like the Wilson, NRTL or Uniquac models (Prausnitz et al., 2000). Alternatively, a purely mathematically motivated modeling approach could be used to correlate states with fluxes or rate coefficients in the sense of black-box modeling. Commonly used model structures include multivariate linear or polynomial models, neural networks, or vector machines among others (Hastie et al., 2003). This way, a *certain type of hybrid (or grey-box) model* (Psichogios and Ungar, 1992; Agarwal, 1997; Olivera, 2004) arises in a natural way by combining first principles models fixed on previous decision levels with an empirical model on the current decision level (Kahrs, Marquardt, 2008; Romijn et al., 2008; Kahrs et al., 2009).

### 3.2 Simultaneous model identification

IMI exploits the natural hierarchy in kinetic models of process systems. All the established approaches to model identification, however, neglect this inherent structure. The so-called simultaneous model identification (SMI) approaches always assume that the model structure is correct and consider only the fully specified model. In particular, the decisions on the balance envelope and the desired spatio-temporal resolution, the selection of the models for the flux expression (*BF*) and the phenomenological coefficients (*BFR*) are specified prior to adjusting the model response to the measured data by some kind of identification method. Since the sub-models are typically not known, suitable model structures are selected by the modeller based on prior knowledge, experience and intuition. Obviously, the complexity of the decision making process is enormous. The number of alternative model structures grows exponentially with the number of decision levels and number of kinetic phenomena occurring simultaneously in the real system.

Any decision on a sub-model will influence the predictive quality of the identified kinetic model. The model predictions are typically biased if the parameter estimation is based on a model containing structural error (Walter, Pronzato, 1997). The theoretically optimal properties of the maximum likelihood approach to parameter estimation (Bard, 1974) are lost, if structural model mismatch is present. More importantly, in case of biased predictions, it is difficult to identify which of the decisions on a certain sub-model contributed most to the error observed.

One way to tackle these problems in simultaneous identification is the enumeration of all the combinations of the candidate sub-model structures for each kinetic phenomenon. Such combinatorial aggregation inevitably results in a large number of model structures. The computational effort for parameter estimation grows very quickly and calls for high performance computing, even in case of spatially lumped models, to tackle the exhaustive search for the best model indicated by the maximum likelihood objective (Wahl et al., 2006). Even if such a brute force approach were adopted, initialization and convergence of the typically strongly nonlinear parameter estimation problems may be difficult since the (typically large number of) parameters of the overall model have to be estimated in one step (Cheng, Yuan, 1997). The lack of robustness of the computational methods may become prohibitive, in particular, in case of spatially distributed process models if they are nonlinear in the parameters (Karalashvili et al., 2011). Appropriate initial values can often not be found to result in reasonable convergence of an iterative parameter estimation algorithm.

### 3.3 Implementation of the identification methods

After outlining the key ideas of the SMI and IMI methods, some discussion on the requirements for the implementation as a prerequisite for their roll-out in practical applications is presented next.

The implementation of SMI is straightforward and can be based on a wealth of existing theoretical and computational tools. Implicitly, SMI assumes a *suitable experiment* and the *correct model structure* to be available. Then, the following steps have to be enacted:

#### SMI procedure

1. Make sure that all the model parameters are identifiable from the measurements (Walter, Pronzato, 1997). If necessary, employ local identifiability methods (Vajda et al, 1987). Select initial parameter values based on a priori knowledge and intuition.
2. Do initial experiments for selected experimental conditions guided by statistical design of experiments (Mason et al., 2003).
3. Estimate the unknown parameters (Bard, 1974; Schittkowski, 2002; Biegler, 2010), most favourably by a maximum likelihood approach to get unbiased estimates, using the available experimental data.
4. Assess the confidence of the estimated parameters and the predictive quality of the model (Bard, 1974; Walter, Pronzato, 1997).
5. Design optimal experiments *for parameter precision* (Walter, Pronzato, 1990; Pukelsheim, 2006; Franceschini, Macchietto, 2008) and run the experiment.
6. Reiterate the sequence of steps 3 to 5 until no improvement in parameter precision can be obtained.

A number of commercial or open-source tools (Buzzi-Ferraris, Manenti, 2009; Balsa-Canto, Banga, 2010) are available, which can be readily applied to reasonably complex models, in particular to models consisting of algebraic and/or ordinary differential equations. Though this procedure is well established, a number of pitfalls may still occur (Buzzi-Ferraris, Manenti, 2009), which render the application of SMI a challenge even under the most favourable assumptions. An analysis of the literature on applications shows that the identification of (bio-) chemical reaction kinetics has been of most interest to date.

If a set  $S$  of candidate model structures  $M_i$  has to be considered because the correct model structure is unknown, the SMI approach as outlined above cannot be applied without modification. If the *correct model structure*  $M_c$  were included in the set of candidate models, the above SMI procedure has to be modified as follows: Steps 1, 3 and 4 have to be carried out for all the candidate models in the

set  $S$ . A decision on the correct model in the set should not be based on the results of step 4, i.e., the model with highest parameter confidence and the best predictive quality should not be selected, because the experiments carried out so far may not allow to distinguish between competing model candidates. An informed decision requires replacing step 5 by step 5', the optimal design of experiments for *model discrimination* (Walter, Pronzato, 1990; Pukelsheim, 2006; Michalik et al., 2010), to determine experiments, which allow distinguishing between the models with highest confidence. The designed experiments are executed, the parameters in the (so far) most appropriate model structure are estimated. Since the optimal design of experiments relies on initial parameters, which may be incorrect, step 3 and 5' have to be reiterated until the confidence in the most appropriate model structure in the candidate set cannot be improved and hence model  $M_c$  has been found. Then, steps 5, 3 and 4 are reiterated to determine the best possible parameters in the correct model structure.

Only little software support is available to the user for an optimal design of experiments for parameter precision (e.g. VPLAN, Körkel et al., 2004) and even less for model discrimination, which is required for a roll-out of the extended SMI procedure. Only few experimental studies have been reported which tackle model identification in the spirit of the extended SMI procedure.

Obviously, if the correct model structure is not known, it cannot be safely assumed that the correct model structure is part of the candidate set  $S$ ; rather, the correct model, often comprising of a combination of many sub-models, is not known. In this likely case, SMI should be replaced by IMI, the strength of which is to find an appropriate model structure composed of many sub-models. IMI comprises the following steps:

### IMI procedure<sup>2</sup>

1. Decide on a balance envelope, on the desired spatio-temporal resolution and on the extensive quantities to be balanced. Develop *model B* (cf. Fig. 2).
2. Decide on the type of measurements necessary to estimate the unknown fluxes in *model B*.
3. Run informative experiments (following, e.g., a space-filling experiment design (Brendel, Marquardt, 2008) and estimate the unknown fluxes  $J_{...}(z, t)$  or  $j_{...}(t)$  as a function of time and space coordinates using the measurements  $x(z, t)$  or  $x(t)$  and Eqs. (1)–(3). Use appropriate regularization techniques to control error amplification in the solution of this inverse problem (Reinsch, 1967; Engl et al., 1996; Huang, 2001).

---

<sup>2</sup> Note, this IMI procedure is not precise, because its details depend on the type of model considered. The presented procedure is abstracted to roughly cover all types of models.

4. Analyse the state/flux data and define a set of candidate flux models, Eqs. (4), (5), with rate coefficient functions  $k_{\dots}(z,t)$  or  $k'_{\dots}(t)$  parameterized in time and space. Fit the rate coefficient functions  $k_{\dots}(z,t)$  or  $k'_{\dots}(t)$  of all candidate models to the state-flux data. Error-in-variables estimation (Britt, Luecke, 1975) should be used for favourable statistical properties, because both, the dependent fluxes as well as the measured states are subject to error. A constant rate coefficient is obviously a reasonable special case of such a parameterization.
5. Form *candidate models*  $BF_i$  constituting balances and (all or only a few promising) candidate flux models. Re-estimate the parameters in the rate coefficient functions  $k_{\dots}(z,t)$  or  $k'_{\dots}(t)$  in all the *candidate models*  $BF_i$  to reduce the unavoidable bias due to error propagation (Bardow, Marquardt, 2004; Karalashvili et al., 2010). Some kind of regularization of the estimation problem is required to enforce uniqueness of the estimation problem and to control error amplification in the estimates (Kirsch, 1996; Engl et al., 1997). Rank order the updated *candidate models*  $BF_i$  with respect to quality of fit using an appropriate statistical measure such as Akaike's information criterion (Akaike, 1973; Burnham, Anderson, 2002) or posterior probabilities (Stewart et al., 1998). In case of constant rate coefficients, continue with step 8 replacing models  $BFR$  by  $BF$ .
6. Analyse the state/rate-coefficient data and define a set of candidate rate coefficient models  $r_{i,j}$ , Eqs. (6), for promising *candidate models*  $BF_i$ . Make sure that the parameters  $q_{i,j}$  in the candidate rate coefficient models  $r_{i,j}$  are identifiable from the state/rate-coefficient data using identifiability analysis (Walter, Pronzato, 1997). Estimate the parameters  $q_{i,j}$  in the rate coefficient models  $r_{i,j}$  by means of an error-in-variables method (Britt, Luecke, 1975).
7. Form the candidate models  $BFR_{i,j}$  by introducing the rate coefficient models  $r_{i,j}$  in the models  $BF_i$ . Re-estimate the parameters  $q_{i,j}$  in the *candidate models*  $BFR_{i,j}$  to remove the unavoidable bias due to error propagation.
8. Design *optimal experiments for model discrimination* using the set of *candidate models*  $BFR_{i,j}$  to identify the most suitable model structure. Execute the design experiments and re-estimate the parameters  $q_{i,j}$  in the *candidate models*  $BFR_{i,j}$  using the available experimental data. Re-iterate this step until the confidence in the most suitable model structure  $BFR_c$  in the candidate set cannot be improved. If no satisfactory model structure can be identified in the set of candidate models, the set has to be revised by revisiting all previous steps.
9. Design *optimal experiments for parameter precision* using model  $BFR_c$ . Run the experiment and estimate the parameters  $q_c$  in model  $BFR_c$ . Re-iterate this step until the confidence in the parameters cannot be improved. If no sat-

isfactory parameter confidence and prediction quality can be achieved, all previous steps have to be revisited.

A successful implementation of the incremental identification approach requires tailored ingredients such as

- high resolution (in-situ and non-invasive) measurement techniques which provide field data of states like species concentrations, temperature or velocities as a function of time and/or space coordinates;
- algorithms for model-free flux estimation by an inversion of the balance equations; a problem, which is closely related to input estimation problems in systems and control engineering (Hirschhorn, 1979) and to inverse problems (in particular inverse source problems) in applied mathematics (Engl et al., 1997);
- algorithms for efficient function estimation comprising an (ideally error-controlled) adaptive discretization of the unknown flux or rate coefficient functions in time and space coordinates (Brendel, Marquardt, 2009) and robust numerical methods for ill-conditioned, large-scale parameter estimation (Hanke, 1995);
- methodologies for the generation, assessment and selection of the most suitable model structures; and
- model-based methods for the optimal design of experiments (Walter, Pronzato, 1990; Pukelsheim, 2006), which should be adapted to the requirements of IMI.

A detailed discussion of all these areas is definitely beyond the scope of this work. Some more detail in the context of IMI has been given by Marquardt (2005). Some aspects are highlighted in the applications of IMI approach described in the following sections, where recent progress is exemplarily reported for selected kinetic modeling problems of chemical process systems. In particular, reaction kinetics modeling, multi-component diffusion in liquids, and energy transport in falling liquid films will be addressed.

## 4 Reaction kinetics

Mechanistic modeling comprising both, the identification of the most likely mechanism and the quantification of the kinetics of a chemical reaction system, is one of the most relevant and still not yet fully satisfactorily solved tasks in process systems modeling (Berger, 2001). More recently, systems biology (Klipp

et al., 2005) has revived this classical problem in chemical engineering to identify mechanisms, stoichiometry and kinetics of metabolic and signal transduction pathways in living systems (Engl et al., 2009). Though this is the very same problem as in process systems modeling, it is more difficult to solve it successfully, because of three complicating facts: (i) there are severe restrictions to in-vivo measurements of metabolite concentrations with sufficient (spatio-temporal) resolution, (ii) the number of metabolites and reaction steps is often very large, and (iii) the qualitative behaviour of living systems changes with time giving rise to variable-structure models.

IMI has been elaborated in theoretical studies for a variety of reaction systems. Bardow and Marquardt (2004) investigate the fundamental properties of IMI for a very simple reaction kinetic problem to elucidate error propagation and to suggest counteractions. Brendel et al. (2006) work out the IMI procedure for homogenous multi-reaction systems comprising any number of irreversible or reversible reactions. These authors investigate which measurements are required to achieve complete identifiability. They show that the method typically scales linearly with the number of reactions because of the decoupling of the identification of the reaction rate models. The method is validated with a realistic simulation study. The computational effort can be reduced by two orders of magnitude compared to an established SMI approach. Michalik et al. (2007) extend IMI to fluid multi-phase reaction systems. These authors show for the first time, how the intrinsic reaction kinetics can be accessed without the usual masking effects due to interfacial mass transfer limitations. The method is illustrated with a simulated two-phase liquid-liquid reaction system of moderate complexity.

More recently, Amrhein et al. (2010) and Bhatt et al. (2010) have suggested an alternative decoupling method for single- and multi-phase multi-reaction systems, which is based on a linear transformation of the reactor model. The transformed model could be used for model identification in the spirit of the SMI procedure. Pros and cons of the decomposition approach of Brendel et al. (2006) and Michalik et al. (2007) and the one of Amrhein et al. (2010) and Bhatt et al. (2010) have been rigorously analysed and illustrated by means of a simulated case study (Bhatt et al., 2012).

Selected features of IMI are elucidated for this important class of identification problems as follows. *IMI.i* refers to step *i* of the IMI procedure worked out in Section 3.3.

## 4.1 Single-phase reaction systems

Reaction kinetic studies of reaction systems are often carried out in continuously or discontinuously operated stirred tank reactors or in differential flow-through reactors where the spatial dependency of concentrations and temperature can be safely neglected. Typically, the evolution of concentrations, temperatures and flow rates is observed over time. The case of homogeneous reactions in a single phase is considered in this section.

**IMI.1-IMI.3: Reaction flux estimation.** The material balances for the mole number  $n_i$  of the  $n_c$  chemical species  $i$  specialize Eqs. (2) and (3) to result in *model B*, i.e.,

$$\frac{dn_i(t)}{dt} = q(t)c_i^{\text{in}}(t) - q(t)c_i(t) + f_i(t), \quad c_i(t)V(t) = n_i(t), \quad i = 1, \dots, n_c. \quad (7)$$

The first two terms on the right hand side refer to the molar flow rates into and out of the reactor with known (or measured) molar flow rate  $q(t)$  and inlet concentrations  $c_i^{\text{in}}(t)$ .<sup>3</sup> The last term represents the unknown reaction flux of species  $i$ , i.e. the molar amount of species  $i$  produced or consumed by all chemical reactions present. The measured concentrations  $c_i(t)$  are converted into the extensive mole numbers  $n_i(t)$  by multiplication with the known (or measured) reactor volume  $V(t)$ . It should be noted that the fluxes enter the balance equations linearly and the equations are decoupled for each species. All reaction fluxes  $f_i(t)$  can thus be estimated individually by numerical differentiation of measured concentration data for each measured species from the material balances. This ill-posed inverse problem can successfully be solved by Tikhonov-Arsenin filtering (Tikhonov, Arsenin, 1977; Mhamdi, Marquardt, 1999) or smoothing splines (Huang, 2001; Bardow, Marquardt, 2004). Regularization parameter choice based on the L-curve (Hansen, O'Leary, 1993) or generalized cross-validation (Golub et al., 1979) has been shown to give reliable estimates.

**IMI.4: Reaction rate models.** The reactions fluxes refer to the total amount of a certain species produced or consumed in a reaction system. Since any chemical species  $i$  participates in more than one reaction  $j$  in a multi-reaction system, the

---

<sup>3</sup> Note that we tacitly assume measurements, which are continuous in time to simplify the presentation. Obviously, real measurements are taken on a grid of discrete times. Hence, the equations may have to be interpreted accordingly.



reaction rates  $r_{ij}$  have to be determined from the reaction fluxes  $f_i$ ,  $i=1,\dots,n_c$ , by solving the (usually non-square) linear system

$$f(t) = V(t)N^T r(t) \quad (8)$$

for  $r(t)$  using an appropriate numerical method. The symbol  $f(t)$  refers to the vector of  $n_c$  reaction fluxes,  $r(t)$  to the vector of reaction rates of the  $n_r$  reactions in the reaction system,  $V(t)$  to the reactor volume and  $N$  to the stoichiometric matrix of appropriate dimension. Often the reaction stoichiometry is unknown; then, target factor analysis (Bonvin, Rippin, 1990) can be used to determine the number of relevant reactions and to test candidate stoichiometries suggested by chemical research. If more than one of the conjectured stoichiometric matrices is found to be consistent with the state/flux data, different estimates of  $r(t)$  are obtained in different scenarios to be followed in parallel in subsequent steps. The concentration/reaction-rate data are analyzed next to suggest a set of candidate reaction rate laws (or purely mathematical relations) which relate each of the reaction rates  $r_j(t)$  with the (possibly  $n_c$ ) concentrations  $c(t)$  according to

$$r_j = m_{j,l}(c, \theta_{j,l}), \quad j = 1, \dots, n_r, \quad l \in \Sigma_j. \quad (9)$$

This model assumes isothermal and isobaric experiments, where the quantities  $q_{j,l}$  are constants. A model selection and discrimination problem has to be solved subsequently for each of the reaction rates  $r_j$  based on the sets of model candidate  $S_j$  because the correct or at least best model structures are not known. These problems are, however, independent of each other. At first, the parameters  $q_{j,l}$  in Eq. (9) are estimated from  $r_j(t)/c(t)$  data by means of nonlinear algebraic regression (Bard, 1974; Walter, Pronzato, 1997). The quality of fit is evaluated by some means to assess whether the conjectured model structures (9) fit the data sufficiently well.

**IMI.5: Reducing the bias and ranking the reaction model candidates.** Eqs. (8) and (9) are now inserted into Eqs. (7) to form a complete reactor model. The parameters in the rate laws (9) are now re-estimated by a suitable dynamic parameter estimation method such as multiple shooting (Lohmann et al., 1992) or successive single shooting (Michalik et al., 2009). Obviously, only the models in the subsets  $S_{j,p}$  of the sets  $S_j$  in Eq. (9) are considered which have been identified to fit the data reasonably well. Very fast convergence is obtained, i.e., often a single iteration is sufficient, because of the very good initial parameter estimates obtained in step IMI.4. This dynamic parameter estimation reduces the bias in the parameter estimates computed in step IMI.4. The model candidates can now be

rank ordered, for example, by Akaike's information criterion (Akaike, 1973) for a first assessment of their relative predictive qualities.

**IMI.6 and IMI.7: Rate coefficient models.** In case of non-isothermal experiments, the quantities  $q_{j,l}$  in the rate models (9) are functions of temperature  $T$ . In this case,  $q_{j,l}$  can be replaced by  $k'_{j,l}$ , which has to be estimated first without specifying a rate coefficient model as in step IMI.6. Then, Eq. (9) is modified and a parameterized rate coefficient model, such as the Arrhenius law,

$$k'_{j,l} = \theta_{j,1} e^{\frac{\theta_{j,2}}{T}}, \quad r_j = k'_{j,l} m_{j,l}(c, \theta_{j,l}), \quad j = 1, \dots, n_r, \quad l \in \Sigma_j. \quad (10)$$

is introduced and the constant parameters  $q_{j,1}$  and  $q_{j,2}$  are estimated from the  $k'_{j,l}(t) / T(t)$  data for every reaction  $j$  (see Brendel (2006) for details).

**IMI.8 and IMI.9: Selecting the best reaction model.** The identification of the reaction rate models may not immediately result in reliable model structures and parameters because of a lack of information content in the experimental data. Iterative improvement with optimally chosen experimental conditions should therefore be employed. Optimal experiments are designed first for model structure discrimination and then, after convergence, for parameter precision to yield the best model  $m_{j,b}(c, q_{j,b})$  contained in the candidate sets  $S_{j,p}$  for all  $j=1, \dots, n_r$ .

**Experimental validation.** The development of the IMI approach solely relied on theoretical considerations which were supported by simulation case studies to validate the method and investigate its properties. An experimental validation of IMI has been carried out (Michalik et al., 2007; Schmidt et al., 2009) for an enzymatic reaction, i.e., the regeneration of  $NAD^+$  to  $NADH$ , a cofactor used in many industrial enzymatic reactions where it is reduced to  $NAD^+$ . The reaction takes place in aqueous solution using formic acid as a proton donor. There are two reactions of interest, the reversible regeneration reaction which forms  $NADH$  and  $CO_2$  as a by-product, and an undesired irreversible decomposition of the product  $NADH$ . The experiments were carried out in a micro-cuvette reactor of 300 ml, where the  $NADH$  concentration was measured with high accuracy and high resolution using UV/Vis spectroscopy at an excitation wavelength of 340 nm. The application of IMI to this industrially relevant problem (Michalik et al. 2007) resulted in a reaction kinetic model with much better predictive quality compared to existing and widely used literature models (Schmidt et al., 2009).

## 4.2 Multi-phase reaction systems

The application of IMI to multi-phase reactions is of great practical interest, because it is extremely difficult to access the intrinsic kinetics of a chemical reaction, which is completely independent of mass transfer effects. Current practice in kinetic modeling of two-phase systems aims at experimental conditions where the chemical reaction is clearly rate-limiting and the effect of the (very fast) mass transfer between the phases can be safely neglected. Obviously, this strategy is quite restrictive and inevitably results in systematic errors in reaction kinetics due to mass transfer contributions. IMI can remedy this long-standing problem in a straightforward manner.

Let us assume isothermal experiments in a stirred tank reactor, which is operated in batch mode (e.g. no material is exchanged with the environment) at isothermal conditions. A liquid-liquid (or liquid-gas) reaction is carried out, where the reaction occurs in one of the phases, say (*a*), only. The experiment is set up such that two well mixed segregated phases (*a*) and (*b*) occur where spatial dependencies of the state variables are negligible. This assumption can easily be implemented by means of appropriate mixing and stabilization of the interface. Concentrations  $c_i^{(a)}(t)$  and  $c_i^{(b)}(t)$  of the relevant species are assumed to be measured (for example by some kind of optical spectroscopy) in both phases. The material balances, specializing Eqs. (2) and (3), read as

$$V^{(a)} \frac{dc_i^{(a)}(t)}{dt} = j_i(t) + f_i(t), \quad V^{(b)} \frac{dc_i^{(b)}(t)}{dt} = -j_i(t), \quad i = 1, \dots, n_c. \quad (11)$$

The volumes  $V^{(a)}$  and  $V^{(b)}$  of both phases are assumed constant and known for the sake of simplicity. The symbols  $j_i(t)$  and  $f_i(t)$  refer to the mass transfer rate of species from phase (*b*) to phase (*a*) and the reaction flux in phase (*a*), respectively.

Steps *IMI.1* to *IMI.3* have to be slightly modified compared to the case of homogenous reaction systems discussed in Section 3.1. In particular, the balance of phase (*b*) (on the right in Eq. (11)) and the measurements of the concentrations  $c_i^{(b)}(t)$  are used to estimate the mass transfer rates  $j_i(t)$  first without specifying a mass transfer model. These estimated functions can be inserted into the balances of phase (*a*) (on the left in Eq. (11)) to estimate the reaction fluxes  $f_i(t)$  without specifying any reaction rate model. The intrinsic reaction kinetics can easily be identified in the subsequent steps *IMI.4* to *IMI.9* from the concentration measurements  $c_i^{(a)}(t)$  and the estimated reaction fluxes  $f_i(t)$ . Obviously, mass transfer models can be identified in the same manner, if the mass transfer rates

and the concentration measurements in both phases  $c_i^{(a)}(t)$  and  $c_i^{(b)}(t)$  are used accordingly.

This basic idea has been worked out in detail by Michalik et al. (2009) and has been evaluated in a simulated case study of a fluid two-phase system. These authors show that the intrinsic reaction kinetics can indeed be identified at high precision. Work on an experimental validation of IMI for reaction kinetic modeling of fluid two-phase systems is in progress.

## 5 Multi-component diffusion in liquids

Despite extensive and lasting research efforts on diffusive transport, there is still a surprising lack of experimentally validated diffusion models, in particular for complex multi-component liquid mixtures (Bird, 2004). This is in stark contrast to the relevance of the quantitative representation of diffusion to support the design of technical equipment. For example, the interplay of multi-component diffusion and chemical reaction determines the selectivity towards the desired product in industrial reactors. In particular, in micro-reactors, where mixing is only due to diffusion because of the laminar flow conditions, the complex mixing and diffusion patterns are decisive for reactor performance (Bothe et al., 2010).

The application of IMI to diffusive mass transport in liquid systems is featured in this section. It is based on a recently introduced Raman diffusion experiment (Bardow et al., 2003, 2006), where the inter-diffusion of two initially layered liquid mixtures is observed by Raman spectroscopy under isothermal conditions. Concentration profiles  $c_i(z,t)$  of all species are measured on a line in the axis of a tailored cuvette at high resolution in time and space. The IMI procedure outlined in Section 3.3 is instantiated for this particular case as follows.

**IMI.1-IMI.3: Diffusive flux estimation.** The diffusion process is assumed to be well-described by a spatially one-dimensional model. The adaption of the general balance equation (1) results in *model B*, a system of mass balance equations for all species  $i$ :

$$\frac{\partial c_i(z,t)}{\partial t} = -\frac{\partial J_i(z,t)}{\partial z}, \quad \left. \frac{\partial c_i(z,t)}{\partial z} \right|_{z=0, z=l} = 0, \quad i = 1, \dots, n_c - 1. \quad (12)$$

The molar concentrations  $c_i(z,t)$  are determined from Raman spectra by means of indirect hard modeling (Alsmeyer et al., 2004, Kriesten et al., 2008) at high accuracy. The  $n_c-1$  independent diffusive fluxes  $J_i(z,t)$  are unknown and have to be inferred from Eqs. (12) by an inversion of each of the evolution equations using

the measured concentration profiles. In particular, the measurements have to be differentiated with respect to time  $t$  first using smoothing splines (Reinsch, 1967) and appropriate regularization (Engl et al., 1997), and the result has to be integrated over the spatial coordinate  $z$  next to render the diffusive fluxes  $J_i(z, t)$ ,  $i=1, \dots, n_c-1$ , without specifying a diffusion model. Such a strategy has been followed for binary and ternary systems by Bardow et al. (2003, 2006). Again, there is only a linear increase in complexity due to the natural decoupling of the multi-component material balances (12).

**IMI.4: Diffusion flux models.** One or more flux models have to be introduced next. The generalized Fick model (or the Maxwell-Stefan model, which is not further considered here) is a suitable choice. In case of binary mixtures, the Fick diffusion coefficient  $D_{1,2}(z, t)$  can be determined at any point in time and space by solving the flux equation

$$J_1(z, t) = -D_{1,2}(z, t) \frac{\partial c_1(z, t)}{\partial z} \quad (13)$$

for  $D_{1,2}(z, t)$ . This strategy does not carry over directly to multi-component mixtures because the diffusive flux is a linear combination of all concentration gradients:

$$J_i(z, t) = - \sum_{j=1}^{n_c-1} D_{i,j}(z, t) \frac{\partial c_j(z, t)}{\partial z}, \quad i = 1, \dots, n_c - 1. \quad (14)$$

Rather, the  $n_c-1$  diffusion coefficients have to be parameterized somehow. For example, some approximating spatio-temporal function could be chosen to formulate a least-squares problem which determines the diffusion coefficients  $D_{i,j}(z, t)$  as function of time and space coordinates. Alternatively, a physically based parameterization (e.g., a diffusion coefficient model) could be chosen to lump IMI.4 and IMI.6 and eliminate IMI.5.

**IMI.5: Reducing the bias.** The *model BF* can be formed by introducing Eqs. (14) into Eqs. (13). The diffusion coefficient functions can be re-estimated using the results of the last step IMI.4 as initial values of the parameter estimation problem to reduce the bias due to error propagation.

**IMI.6 and IMI.7: Diffusion coefficient models.** Diffusion coefficient models can now be chosen to correlate the estimated diffusion coefficient data with the measured concentrations:

$$D_{i,j} = m_{i,j,l}(c, \theta_{j,l}), \quad i, j = 1, \dots, n_c - 1, \quad l \in \Sigma_{i,j}. \quad (15)$$

Again, a model selection problem has to be solved. The parameters  $q_{j,l}$  are identified by error-in-variables estimation (Britt, Luecke, 1975). The bias can be removed by inserting Eq. (15) into Eqs. (14) and the result into Eq. (12) and re-estimating the parameters. The models can be ranked with respect to model quality by some statistical measure (Burnham, Anderson, 2002; Stewart et al., 1998).

**IMI.8 and IMI.9: Selecting the best diffusion model.** To remedy the possible lack of information content in the experimental data an iterative improvement with optimally chosen experimental conditions should finally be employed to yield the best diffusion models  $D_{i,j}$ .

**Experimental validation.** The suggested strategy has been validated in a number of experimental studies including the determination of binary and ternary Fick diffusion coefficients with a very low number of Raman experiments (Bardow et al., 2003, 2006) and the identification of the full concentration dependency of the binary Fick diffusion coefficient by means of a single Raman inter-diffusion experiment (Bardow et al., 2005) and two additional NMR self-diffusion experiments at infinite dilution to improve accuracy (Kriesten et al., 2009).

## 6 Energy transport in falling liquid films

The applicability of IMI to relevant and challenging problems has been demonstrated in the two previous sections. Still, the complexity tackled has been moderate, since three-dimensional (3D), transient transport and reaction problems in complex spatial geometries have not yet been treated. Such problems are relevant not only in chemical process systems, but in many other areas of science and engineering. As a first step towards the application of IMI to general 3D transient transport and reaction problems the identification of a transport coefficient function in the energy equation of a model of a wavy falling film has been chosen (Karalashvili et al., 2008, 2011).

Falling liquid films are widely used in chemical engineering, e.g., to implement coolers, evaporators, absorbers or chemical reactors, where the wavy surface patterns are exploited to intensify heat and mass transfer between the liquid film and the surrounding gas. Even the dynamics of heated falling films of a single chemical species is complex and has been the subject of intensive research (e.g., Trevelyan et al., 2007; Meza, Balakotaiah, 2008). Direct numerical

simulation of the free-surface, mixed initial-boundary problem involving the continuity, the momentum and the energy equations is very involved and has not yet been reported to the author's knowledge. Even if it were possible, the computational complexity would prevent its application for the design of technical equipment. As an alternative, Wilke (1962) suggested a long time ago to approximate the complex spatial domain of the wavy liquid film by a flat-film geometry and to introduce a so-called *effective transport coefficient* which has to account for the wave-induced backmixing present in the wavy film (Adomeit, Renz, 2000). Yet, there are no accepted and reasonably general models available, which correlate the effective transport coefficient with the velocity and temperature fields in the falling film. The IMI procedure seems to be a promising starting point to tackle this long-standing problem by the sequence of steps outlined in Section 3.3 as follows.

**IMI.1-IMI.3: Diffusive energy flux estimation.** The energy transport in a 3D, transient, flat falling film can be represented by the energy equation, which can be reformulated for incompressible fluids (with constant density  $\rho$ ) to result in

$$\rho \frac{\partial u}{\partial t} = -\rho w \cdot \nabla u - \nabla \cdot J_u \quad (16)$$

with appropriate initial and boundary conditions. The velocity field  $w(z,t)$  is assumed to be known (either measured or computed from a possibly approximate solution of the Navier-Stokes equations), while the internal energy  $u(z,t)$  (or rather the temperature  $T(z,t)$ ) is assumed to be measured at reasonable spatio-temporal resolution. This *model B* can be refined by decomposing the diffusive energy flux  $J_u(z,t)$  into a known molecular and an unknown *wave-induced* term. This reformulation results finally in

$$\frac{\partial T(z,t)}{\partial t} + w(z,t) \cdot \nabla T(z,t) - \nabla \cdot [a_{mol}(z,t) \nabla T(z,t)] = F_{wavy}(z,t) \quad (17)$$

with the known molecular transport coefficient  $a_{mol}(z,t)$  and the unknown wavy contribution to the energy flux  $F_{wavy}(z,t)$ . This flux contribution can be reconstructed from temperature field data by solving a source inverse problem, which is linear in the unknown  $F_{wavy}(z,t)$  by an appropriate regularized numerical method (Karatashvili et al., 2008).

**IMI.4: Wavy energy flux model.** A reasonable model for the wavy contribution to the energy flux is motivated by Fourier's law. Hence, the flux  $F_{wavy}(z,t)$  in Eq. (17) can be related to wavy transport coefficient  $a_{wavy}(z,t)$  by the ansatz

$$F_{wavy}(z, t) = -\nabla \cdot J_{u, wavy}(z, t) = -\nabla \cdot (a_{wavy}(z, t) \nabla T(z, t)). \quad (18)$$

Note that the effective transport coefficient is defined as the sum of the molecular and the wavy transport coefficients, i.e.,  $a_{eff} = a_{mol} + a_{wavy}$ . In order to estimate  $a_{wavy}(z, t)$ , a (nonlinear) coefficient inverse problem in the spatial domain has to be solved for any point in time  $t$  (Karalashvili et al., 2008).

**IMI.5: Reducing the bias.** The *model BF* is formed by introducing Eq. (18) into Eq. (17). The resulting equation is used to re-estimate the wavy coefficient  $a_{wavy}(z, t)$ , starting from the estimate in step *IMI.4* as initial values (Karalashvili et al., 2011).

**IMI.6 and IMI.7: Models for the wavy energy transport coefficient.** A set of algebraic models is introduced to parameterize the transport coefficients in time and space by an appropriate model structure:

$$a_{wavy} = m_{wavy, l}(z, t, \theta_{j, l}), \quad l \in \Sigma_j. \quad (19)$$

This set is the starting point for the identification of a suitable parametric model, which properly relates the transport coefficient with velocity and temperature and possibly their gradients. The bias can again be removed by first inserting Eq. (19) into Eqs. (18) and the result into Eq. (17) and next re-estimating the parameters prior to a ranking of the models with respect to model quality (Karalashvili et al., 2011).

**IMI.8 and IMI.9: Selecting the best transport coefficient model.** Optimal design of experiments should finally be employed to obtain most informative measurements to finally identify the best model for  $a_{wavy}(z, t)$  (Karalashvili, Marquardt, 2010).

**Experimental validation** has not yet been possible. For one, the development of this variant of IMI has not yet been completed. Furthermore, high-resolution measurements of film thickness, temperature and velocity fields are mandatory. Optical techniques are under investigation in collaborating research groups (Schagen et al., 2006).



## 7 Concluding discussion

The exemplary applications of IMI as part of the MEXA work process section not only demonstrate its versatility but also its distinct advantages compared to established SMI methods (Bardow and Marquardt, 2004).

### 7.1 Comparing IMI to SMI

In contrast to SMI, the IMI approach explicitly accounts for the fact that often an appropriate structure of one or more sub-models in a complex process systems model is uncertain. The selection of the most suitable sub-model structure has to be considered an integral part of the model identification process. Since model identification cannot be reduced to estimating the parameters from most informative experiments in a given, identifiable model structure, the model (structure) identification process has to be fully transparent to the modeller. Partial prior knowledge regarding model structure can easily be incorporated. Missing sub-models are derived either from experimental or from inferred input-output data in the previous estimation step supported by theoretical investigations on a finer (often the molecular) scale. Any decision on the model structure relates to a single physico-chemical phenomenon and thus reduces ambiguity. Identifiability can be assessed more easily on the level of the sub-model. This way, the IMI strategy supports the discovery of novel model structures which are consistent with the available experimental data.

The decomposition strategy of IMI is also very favourable from a computational perspective. It drastically reduces computational load, because it breaks the curse of dimensionality due to the combinatorial nature of the decision making problem related to sub-model selection. IMI avoids this problem, because the decision making is integrated into the decomposition strategy and systematically exploits knowledge acquired during the previous identification steps. Furthermore, the computational effort is reduced, because the solution of a *strongly nonlinear inverse problem* involving (partial) differential-algebraic equations is replaced by a sequence of less complex, often *linear inverse problems* and a few *algebraic regression problems*. This divide-and-conquer approach also improves the robustness of the numerical algorithms and their sensitivity towards the choice of initial estimates. Last but not least, the decomposition strategy facilitates quasi-global parameter estimation in those cases, where all but the last nonlinear regression problem are convex. A general quasi-global deterministic solution strategy is worked out by Michalik et al. (2009) for identification problems involving differential-algebraic problems.

The computational advantages of IMI become decisive in case of the identification of complex 3D transport and reaction models on complex spatial domains. Our case studies indicate that SMI is computationally often intractable, while IMI renders the estimation problems feasible or at least reduces the load by orders of magnitude. Identifiability analysis and optimal design of experiments are key to success in case of 3D transport and reaction problems, because sufficient excitation in time and space can typically not be achieved intuitively.

Error propagation is unavoidable in IMI, because any estimation error will impair the estimation quality in the following steps. The resulting bias can, however, be easily removed by a final correction step, where a parameter estimation problem is solved for the best aggregated model(s) using very good initial parameter values. Convergence is typically achieved in one or very few iterations.

Both, IMI and SMI are not successful, if the information content of the measurements is insufficient. However, identifiability problems can be discovered and remedied more easily in IMI compared to SMI. Then, either the model has to be simplified (to result in less unknown model parameters) or additional sensors have to be installed in the experiment.

## 7.2 Previous work related to IMI

IMI is not the first multi-step approach to model identification. Similar ideas have been employed rather intuitively before in (bio-)chemical engineering. The sequence of flux estimation and parameter regression is, e.g., commonly employed in reaction kinetics as the so-called differential method (Kittrell, 1970; Hosten, 1979; Froment, Bischof, 1990). Markus et al. (1981) seem to be the first suggesting a simple version of IMI to the identification of enzyme kinetics models. Bastin and Dochain (1990) have introduced model-free reaction flux estimation as part of a state estimation strategy with applications to bioreactors. More recently, a two-step approach has been applied for the hybrid modeling of fermentation processes (Tholudur, Ramirez, 1999; van Lith et al., 2002), where reaction fluxes are estimated first from measured data and neural networks or fuzzy models are employed to correlate the fluxes with the measurements. The crystal growth rate in mixed-suspension crystallization has been estimated directly from the population balance equations (Mahoney et al., 2002).

The idea has not only been around in the chemical engineering community. For example, Timmer et al. (2000) and Voss et al. (2003) use the two-step approach of flux estimation and rate law fitting in the modeling of nonlinear electrical circuits. Ramsay and co-workers used a similar method, called functional data analysis, in quantitative psychology to model lip motion (Ramsay, 1996)

and handwriting (Ramsay, 2000), and in production planning (Ramsay, Ramsay, 2002). These diverse applications and our own experience lead us to the expectation that IMI can be rolled out and tailored to many domains in engineering and the sciences.

### 7.3 Useful models at minimal effort

IMI is considered an integral part of the MEXA methodology. Our experience in a wide area of applications shows that a sensible integration of modeling and experimentation is indispensable if the mathematical model is supposed to extrapolate with adequate accuracy well beyond the region where model identification has been carried out. Such good extrapolation provides at least an indication that the physico-chemical mechanisms underlying the observed system behavior have been captured by the model to a certain extent.

A coordinated design of the model structure and the experiment as advocated in the MEXA work process is most appropriate for several reasons (cf. Bard, 1974; Iyengar and Rao, 1983; Kittrell, 1990; Beck, Woodbury, 1998). On the one hand, an overly detailed model is often not identifiable even if perfect measurements of all the state variables were available (cf. Quaiser and Mönnigmann (2009) for an example from systems biology). Hence, any model should only cover a level of detail, which facilitates an experimental investigation of model validity. On the other hand, an overly simplified model does often not reflect real behaviour satisfactorily. For example, equilibrium tray models in distillation assume phase equilibrium rather than accounting for the mass transfer resistance between the liquid and vapour phases. Though this model is still widely used in industrial practice, it has been shown to be inconsistent with basic physical principles, since it does not reflect the cross-effects of multi-component diffusion (Taylor, Krishna, 1993). Such a coordinated design of experiment and models is closely related to the requirement of refining a model only based on experimental evidence (Markus et al., 1981). In particular, if a model is able to predict the accessible observations on the associated real system sufficiently well, its further refinement cannot be justified because it reduces the level of confidence in the model.

The identification of *useful models at minimal effort* requires a multi-disciplinary team effort. Experts in high-resolution measurement techniques, in the application domain of interest, in numerical analysis and in modeling methodologies have to join forces to leverage the very high effort of model identification. Best-practices and suitable software environments, tailored to a certain application, such as reaction kinetics identification seem to be indispensable to roll out the MEXA framework into routine application.

## 8 Acknowledgements

This work has been carried out as part of CRC 540 “Model-based Experimental Analysis of Fluid Multi-Phase Reactive Systems”, which has been funded by the German Research Foundation (DFG) from 1999 to 2009. The substantial financial support of DFG is gratefully acknowledged. Furthermore, the contributions of the CRC 540 team, in particular however of A. Bardow, M. Brendel, M. Karalashvili, C. Michalik and A. Mhamdi are appreciated.

## References

- Adomeit, P. & Renz, U. (2000). Hydrodynamics of three-dimensional waves in laminar falling films. *International Journal of Multiphase Flow* 26(7). 1183–1208.
- Agarwal, M. (1997). Combining neural and conventional paradigms for modeling, prediction and control. *International Journal of Systems Science* 28. 65–81.
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In: Petrov, B. N. & Csaki, F. (eds.). *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado. 267–281.
- Alsmeyer, F., Koß, H.-J., & Marquardt, W. (2004). Indirect spectral hard modeling for the analysis of reactive and interacting mixtures. *Journal of Applied Spectroscopy* 58(8). 975–985.
- Amrhein, M., Bhatt, N., Srinivasan, B., & Bonvin, D. (2010). Extents of reaction and flow for homogeneous reaction systems with inlet and outlet streams. *AIChE Journal* 56(11), 2873–2886.
- Asprey, S. P. & Macchietto, S. (2000). Statistical tools in optimal model building. *Computers and Chemical Engineering* 24. 1261–1267.
- Balsa-Canto, E. & Banga, J. R. (2010). AMIGO: A model identification toolbox based on global optimization and its applications in biosystems. *11th IFAC Symposium on Computer Applications in Biotechnology*. Leuven, Belgium.
- Bard, Y. (1974). *Nonlinear Parameter Estimation*. New York: Academic Press.
- Bardow, A., Marquardt, W., Göke, V., Koß, H. J., & Lucas, K. (2003). Model-based measurement of diffusion using Raman spectroscopy. *AIChE Journal* 49(2). 323–334.
- Bardow, A. & Marquardt, W. (2004). Incremental and simultaneous identification of reaction kinetics: methods and comparison. *Chemical Engineering Science* 59(13). 2673–2684.
- Bardow, A., Göke, V., Koß, H.-J., Lucas, K., & Marquardt, W. (2005). Concentration-dependent diffusion coefficients from a single experiment using model-based Raman spectroscopy. *Fluid Phase Equilibria* 228–229. 357–366.
- Bardow, A., Göke, V., Koß, H. J., & Marquardt, W. (2006). Ternary diffusivities by model-based analysis of Raman spectroscopy measurements. *AIChE Journal* 52(12). 4004–4015.
- Bardow, A., Bischof, C., Bücker, M., Dietze, G., Kneer, R., Leefken, A., Marquardt, W., Renz, U., & Slusanschi, E. (2008). Sensitivity-based analysis of the k- $\epsilon$ -Model for the turbulent flow between two plates. *Chemical Engineering Science*, 63. 4763–4776.

- Bardow, A. & Marquardt, W. (2009). Identification methods for reaction kinetics and transport. In: Floudas, C. A., & Pardalos, P. M. (eds.): *Encyclopedia of Optimization*, 2<sup>nd</sup> Edition. Berlin: Springer. 1549–1556.
- Bastin, G. & Dochain, D. (1990). *On-line Estimation and Adaptive Control of Bioreactors*. Amsterdam: Elsevier.
- Beck, J. V. & Woodbury, K. A. (1998). Inverse problems and parameter estimation: integration of measurements and analysis. *Measurement Science and Technology* 9(6). 839–847.
- Bhatt, N., Amrhein, M., & Bonvin, D. (2010). Extents of reaction, mass transfer and flow for gas-liquid reaction systems. *Industrial & Engineering Chemical Research* 49(17). 7704–7717.
- Bhatt, N., Kerimoglu, N., Amrhein, M., Marquardt, W., & Bonvin, D. (2012). Incremental model identification for reaction systems – A comparison of rate-based and extent-based approaches. *Chemical Engineering Science* 83. 24–38.
- Biegler, L. T. (2010). *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*. Philadelphia: SIAM.
- Bird, R. B., Stewart, W. E., & Lightfoot, E. N. (2002). *Transport Phenomena*. 2<sup>nd</sup> Edition. New York: Wiley.
- Bird, R. B. (2004). Five decades of transport phenomena. *AIChE Journal* 50(2), 273–287.
- Bothe, D., Lojewski, A., & Warnecke, H.-J. (2010). Computational analysis of an instantaneous irreversible reaction in a T-microreactor. *AIChE Journal* 56(6). 1406–1415.
- Brendel, M. (2006). Incremental Identification of Complex Reaction Systems. *Fortschritt-Berichte VDI* 864. Erlensee: VDI-Verlag.
- Brendel, M. & Marquardt, W. (2008). Experimental design for the identification of hybrid reaction models from transient data. *Chemical Engineering Journal* 141. 264–277.
- Brendel, M. & Marquardt, W. (2009). An algorithm for multivariate function estimation based on hierarchically refined sparse grids. *Computing and Visualization in Science* 12(4). 137–153.
- Britt, H. I. & Luecke, R. H. (1975). Parameter estimation with error in observables. *American Journal of Physics* 43(4). 372.
- Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2<sup>nd</sup> Edition. New York: Springer.
- Buzzi-Ferraris, G. & Manenti, F. (2009). Kinetic models analysis. *Chemical Engineering Science* 64(5). 1061–1074.
- Cheng, Z. M. & Yuan, W. K. (1997). Initial estimation of heat transfer and kinetic parameters of a wall-cooled fixed-bed reactor. *Computers and Chemical Engineering* 21(5). 511–519.
- Engl, H. W., Hanke, M., & Neubauer, A. (1996). *Regularization of Inverse Problems*. Dordrecht: Kluwer.
- Engl, H. W., Flamm, C., Kügler, P., Lu, J., Müller, S., & Schuster, P. (2009). Inverse problems in systems biology. *Inverse Problems* 25. 123014.
- Franceschini, G. & Macchietto, S. (2008). Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science* 63(19). 4846–4872.
- Froment, G. F. & Bischoff, K. B. (1990). *Chemical Reactor Analysis and Design*. New York: Wiley.
- Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics* 21(2). 215–223.
- Hanke, M. (1995). *Conjugate Gradient Type Methods for Ill-Posed Problems*. Essex: Longman Scientific and Technical.

- Hansen, P. C. & O'Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing* 14(6). 1487–1503.
- Hastie, T., Tibshirani, R., & J. Friedman (2003). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Higham, D. J. (2008). Modeling and simulating chemical reactions. *SIAM Review* 50, 347–368.
- Hirschorn, R. M. (1979). Invertibility of nonlinear control systems. *SIAM Journal on Control and Optimization* 17, 289–297.
- Hosten, L. H. (1979). A comparative study of short cut procedures for parameter estimation in differential equations. *Computers and Chemical Engineering* 3, 117–126.
- Huang, C. (2001). Boundary corrected cubic smoothing splines. *Journal of Statistical and Computational Simulation* 70, 107–121.
- Iyengar, S. S. & Rao, M. S. (1983). Statistical techniques in modeling of complex systems – single and multiresponse models. *IEEE Transactions on Systems, Man, and Cybernetics* 13(2), 175–189.
- Kahrs, O. & Marquardt, W. (2008). Incremental identification of hybrid process models. *Computers and Chemical Engineering* 32(4–5), 694–705.
- Kahrs, O., Brendel, M., Michalik, C., & Marquardt, W. (2009). Incremental identification of hybrid models of process systems. In: Hof, P. M. J. van den, Scherer, C., & Heuberger, P. S. C. (eds.). *Model-Based Control*. Dordrecht: Springer. 185–202.
- Karalashvili, M., Groß, S., Mhamdi, A., Reusken, A., & Marquardt, W. (2008). Incremental identification of transport coefficients in convection-diffusion systems. *SIAM Journal on Scientific Computing* 30(6), 3249–3269.
- Karalashvili, M. & Marquardt, W. (2010). Incremental identification of transport models in falling films. *International Symposium on Recent Advances in Chemical Engineering, IIT Madras, December 2010*.
- Karalashvili, M., Groß, S., Marquardt, W., Mhamdi, A., & Reusken, A. (2011). Identification of transport coefficient models in convection-diffusion equations. *SIAM Journal on Scientific Computing* 33(1), 303–327.
- Kirsch, A. (1996). *An Introduction to the Mathematical Theorie of Inverse Problems*. New York: Springer.
- Kittrell, J. R. (1970). Mathematical modeling of chemical reactions. *Advances in Chemical Engineering*, 8, 97–183.
- Klipp, E., Herwig, R., Kowald, A., Wierling, C., & Lehrach, H. (2005). *Systems Biology in Practice. Concepts, Implementation, and Application*. Weinheim: Wiley.
- Körkel, S., Kostina, E., Bock, H. G., & Schlöder, J. P. (2004). Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes. *Optimization Methods and Software* 19(3–4), 327–338.
- Kriesten, E., Alsmeyer, F., Bardow, A., & Marquardt, W. (2008). Fully automated indirect hard modeling of mixture spectra. *Chemometrics and Intelligent Laboratory Systems* 91, 181–193.
- Kriesten, E., Voda, M. A., Bardow, A., Göke, V., Casanova, R., Blümich, B., Koß, H.-J., & Marquardt, W. (2009). Direct determination of the concentration dependence of diffusivities using combined model-based Raman and NMR experiments. *Fluid Phase Equilibria* 277, 96–106.
- Lith, P. F. van, Betlem, B. H. L., & Roffel, B. (2002). A structured modeling approach for dynamic hybrid fuzzy-first principles models. *Journal of Process Control* 12(5), 605–615.

- Lohmann, T., Bock, H. G., & Schlöder, J. P. (1992). Numerical methods for parameter estimation and optimal experiment design in chemical reaction systems. *Industrial and Engineering Chemical Research* 31(1). 54–57.
- Mahoney, A. W., Doyle, F. J., & Ramkrishna, D. (2002). Inverse problems in population balances: Growth and nucleation from dynamic data. *AIChE Journal* 48(5). 981–990.
- Markus, M., Plessner, T., & Kohlmeier, M. (1981). Analysis of progress curves in enzyme kinetics – bias and convergent set in the differential and in the integral method. *Journal of Biochemical and Biophysical Methods* 4(2). 81–90.
- Marquardt, W. (1995). Towards a process modeling methodology. In: Berber, R. (ed.). *Methods of Model-Based Control*, NATO-ASI Ser. E, Applied Sciences. Dordrecht: Kluwer. 3–41.
- Marquardt, W., Wedel, L. von, & Bayer, B. (2000). Perspectives on lifecycle process modeling. *AIChE Symposium Series* 26(323). 192–214.
- Marquardt, W. (2005). Model-based experimental analysis of kinetic phenomena in multi-phase reactive systems. *Chemical Engineering Research and Design* 83(A6). 561–573.
- Mason, R. L., Gunst, R. F., & Hess, J. L. (2003). *Statistical Design and Analysis of Experiments – With Applications to Engineering and Science*. 2<sup>nd</sup> Edition. New York: Wiley.
- Michalik, C., Schmidt, T., Zavrel, M., Ansorge-Schumacher, M., Spieß, A., & Marquardt, W. (2007). Application of the incremental identification method to the formate oxidation using formate dehydrogenase. *Chemical Engineering Science* 62(3). 5592–5597.
- Michalik, C., Hannemann, R., & Marquardt, W. (2009). Incremental single shooting – a robust method for the estimation of parameters in dynamical systems. *Computers and Chemical Engineering* 33. 1298–1305.
- Michalik, C., Brendel, M., & Marquardt, W. (2009). Incremental identification of fluid multi-phase reaction systems. *AIChE Journal* 55(4). 1009–1022.
- Michalik, C., Chachuat, B., & Marquardt, W. (2009). Incremental global parameter estimation in dynamical systems. *Industrial and Engineering Chemistry Research* 48. 5489–5497.
- Meza, C. E., & Balakotaiah, V. (2008). Modeling and experimental studies of large amplitude waves on vertically falling films. *Chemical Engineering Science* 63. 4704–4734.
- Oliveira, R. (2004). Combining first principles modeling and artificial neural networks: a general framework. *Computers and Chemical Engineering* 28. 755–766.
- Pope, S. B. (2000). *Turbulent Flows*. Cambridge: Cambridge Univ. Press.
- Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.
- Prausnitz, J. M., Lichtenthaler, R. N., & Gomes de Azevedo, E. (2000). *Molecular Thermodynamics of Fluid-Phase Equilibria*. 3rd Edition. New Jersey: Prentice Hall.
- Psichogios, D. C. & Ungar, L. H. (1992). A hybrid neural network – first principles approach to process modeling. *AIChE Journal* 38. 1499–1511.
- Pukelsheim, F. (2006). *Optimal Design of Experiments*. Philadelphia: SIAM.
- Quaiser, T. & Mönnigmann, M. (2009). Systematic identifiability testing for unambiguous mechanistic modeling – application to JAK-STAT, MAP kinase, and NF-kappa B signaling pathway models. *BMC Systems Biology* 3. 50.
- Quaiser, T., Dittrich, A., Schaper, F., & Mönnigmann, M. (2011). A simple workflow for biologically inspired model reduction – application to early JAK-STAT signaling. *BMC Systems Biology* 5. 30.
- Ramsay, J. O., Munhall, K. G., Gracco, V. L., & Ostry, D. J. (1996). Functional data analyses of lip motion. *Journal of the Acoustical Society of America* 99(6). 3718–3727.
- Ramsay, J. O. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association* 95(449). 9–15.



- Ramsay, J. O. & Ramsey, J. B. (2002). Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics* 107(1–2). 327–344.
- Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik* 10. 177–183.
- Romijn, R., Özkan, L., Weiland, S., Ludlage, J., & Marquardt, W. (2008). A grey-box modeling approach for the reduction of nonlinear systems. *Journal of Process Control* 18(9). 906–914.
- Schagen, A., Modigell, M., Dietze, G., & Kneer, R. (2006). Simultaneous measurement of local film thickness and temperature distribution in wavy liquid films using a luminescence technique. *International Journal of Heat and Mass Transfer* 49(25–26). 5049–5061.
- Schittkowski, K. (2002). Numerical Data Fitting in Dynamical Systems: A Practical Introduction with Applications and Software. Dordrecht: Kluwer.
- Schmidt, T., Michalik, C., Zavrel, M., Spieß, A., Marquardt, W., & Ansorge-Schumacher, M. (2009). Mechanistic model for prediction of formate dehydrogenase kinetics under industrially relevant conditions. *Biotechnology Progress* 26. 73–78.
- Slattery, J. C. (1999). *Advanced Transport Phenomena*. Cambridge: Cambridge University Press.
- Stewart, W. E., Shon, Y., & Box, G. E. P. (1998). Discrimination and Goodness of Fit of Multiresponse Mechanistic Models. *American Institute of Chemical Engineers Journal* 44(6). 1404–1412.
- Takamatsu (1983). The nature and role of process systems engineering. *Computers and Chemical Engineering* 7(4). 203–218.
- Taylor, R. & Krishna, R. (1993). *Multicomponent Mass Transfer*. New York: Wiley.
- Tholudur, A. & Ramirez, W. F. (1999). Neural-network modeling and optimization of induced foreign protein production. *AIChE Journal* 45(8). 1660–1670.
- Tikhonov, A. N. & Arsenin, V. Y. (1977). *Solution of Ill-Posed Problems*. Washington: V. H. Winston and Sons.
- Timmer, J., Rust, H., Horbelt, W., & Voss, H. U. (2000) Parametric, nonparametric and parametric modeling of a chaotic circuit time series. *Physics Letters A* 274(3–4). 123–134.
- Trevelyan, P. M. J., Scheid, B., Ruyer-Quil, C., & Kalliadasis, S. (2007). Heated falling films. *Journal of Fluid Mechanics* 592. 295–334.
- Vajda, S., Rabitz, H., Walter, E., & Lecourtier, Y. (1989). Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications* 83. 191–219.
- Verheijen, P. J. T. (2003). Model selection: An overview of practices in chemical engineering. In: Asprey, S. P., & Macchietto, S. (eds.). *Dynamic Model Development: Methods, Theory and Applications*. Amsterdam: Elsevier. 85–104.
- Voss, H. U., Rust, H., Horbelt, W., & Timmer, J. (2003). A combined approach for the identification of continuous non-linear systems. *International Journal of Adaptive Control and Signal Processing* 17(5). 335–352.
- Wahl, S. A., Haunschild, M. D., Oldiges, M., & Wiechert, W. (2006). Unravelling the regulatory structure of biochemical networks using stimulus response experiments and large-scale model selection. *IEE Proceedings – Systems Biology* 153(4). 275–285.
- Walter, E. & Pronzato, L. (1990). Qualitative and quantitative experiment design for phenomenological models – a survey. *Automatica* 26(2). 195–213.
- Walter, E. & Pronzato, L. (1997). *Identification of Parametric Models from Experimental Data*. Berlin: Springer.
- Wilke, W. (1962). Wärmeübergang an Rieselfilmen. *VDI-Forschungs-Heft* 490. Düsseldorf: VDI-Verlag.



**Prof. Dr.-Ing. Wolfgang Marquardt**

RWTH Aachen University

AVT-Process Systems Engineering (EPT)

52056 Aachen

Germany

wolfgang.marquardt@avt.rwth-aachen.de



Robin Findlay Hendry

# Kinetics, Models, and Mechanism

Commentary on Wolfgang Marquardt

## 1 Chemical Engineering vs. Chemistry

Wolfgang Marquardt describes chemical engineering as ‘an engineering science that focuses on the foundations of any kind of transformation of matter in order to change its molecular or morphological constitution’ (2013, 187). This makes it sound close to the ‘pure science’ of chemistry, and it certainly is, but I would like to start by highlighting some differences between the two kinds of models: in their scope, in the direction of their representational fit, and in the practical and epistemic interests that constrain their construction.

Starting with scope, chemical engineering models often represent complete production processes, including the logistics of the supply of reactants, their preparation, their reaction, and the purification and distribution of the products (Marquardt 2013, 188; van Brakel 2000, Chapter 7; van Brakel 2011, 533). The processes typically studied by chemistry form only part of the overall production process. The basic modeling strategy for dealing with the complexity that comes with broad scope is to break a multi-stage and multi-scale process down into its components (‘unit operations’ such as mixing, flow, chemical transformation and separation), so that the process as a whole is considered as a series of interacting systems. So chemical processes are there at the heart of the production process as modelled by the chemical engineer, and are themselves broken down further, into the basic kinds of chemical change such as oxidation, reduction or polymerisation (van Brakel 2011, 535–7). And as we shall see later, it is crucial to the understanding of reaction kinetics that basic chemical changes themselves are understood to consist of a series of basic kinds of step at the molecular level.

Turning next to direction of fit, one might think of a ‘pure’ scientific model as an abstract mathematical object, which is developed as a representation of some part of the world. The model is amended to fit the world, not the other way round. In contrast, engineering disciplines seek to change the world, not just to understand it. Process development in chemical engineering involves not only the construction and refinement of a mathematical model, but the construction and refinement of a concrete model production process, which is then scaled up to a full production process. The concrete model is designed in the light of the abstract model, while the abstract model is refined in the light of the behaviour of the con-

crete model, and so on. Thus according to Jaap van Brakel, model and reality are mutually attuned in a two-way, rather than a one-way process (van Brakel 2000, Chapter 7; 2011, 545). The interaction between model and experimental system can be complex: van Brakel attributes the growth of *ab initio* design in chemical engineering to the tailoring of chemical processes so that they realise ‘idealised circumstances, circumstances for which the initial and boundary conditions are manageable in such a way that the increasing power of computational methods can be exploited’ (van Brakel 2011, 534). In other words, computational resources constrain the kinds of experimental system that get built. It is possible to overstate the contrast between pure and engineering science, however. Experimental investigation typically involves some kind of material construction: the development of devices that reliably behave in certain ways, displaying new kinds of behaviour (Hacking 1983, Chapter 13), and, given that theoretical understanding evolves in tandem with the device, the process looks similar to van Brakel’s interactive account of engineering science. This, in fact, is just how Jed Buchwald describes the discovery of electric waves in his scientific biography of Heinrich Hertz (Buchwald 1994).

Lastly there are the interests that govern model construction: Marquardt describes the aim of modeling as the construction of ‘useful models at minimum engineering effort’ (Marquardt 2013, 189). This suggests a pragmatic trade-off between, on the one hand, quantitative accuracy (good enough for the practical purposes at hand), and qualitative understanding, versus computational and experimental effort on the other hand. In a commercial environment this pragmatic choice will also involve economic and even environmental considerations (van Brakel 2011, 534).

## 2 The MEXA Methodology

Marquardt presents his ‘model-based experimental analysis’ (or MEXA for short), which is a detailed method for constructing and refining kinetic models of chemical reactions. After a pilot experiment on the relevant reaction, ‘[e]xperimental evidence and the available a-priori knowledge are used to build a first structured mathematical model of the experiment’ (Marquardt 2013, 190). The experimental system is simulated so as to provide some (fallible) insight into its behaviour in non-actual situations, and the experiment itself is then performed under ‘a telling set of experimental conditions’ (2013, 190). This process provides a mixture of qualitative and quantitative information, which acts as a constraint on the refinement of the model. The improved model is then used to refine the

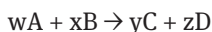
experimental set-up further. The process is repeated until consistency between model and experiment is achieved at the required level of accuracy. As Marquardt puts it, citing Karl Popper's book *The Logic of Scientific Discovery* (Popper 1959): 'The investigations should ideally only be terminated if the model can not be falsified by any conceivable experiment' (2013, 190–191). Marquardt characterises his approach to model identification as *incremental*, in the sense that different aspects of the model's structure are identified step by step, exploiting the 'natural hierarchy in kinetic models of process systems' (2013, 196). He carefully distinguishes this approach from what he calls 'simultaneous model identification' (2013, 196), in which decisions are made simultaneously on the various structural features and parameters that identify the model.

I would like to pick up on two specific features of MEXA mentioned by Marquardt. The first is the pursuit of the method within a framework of *a priori* constraints. This sounds curious: the efficacy of an iterative method like MEXA is highly dependent on the choice of its starting point. If the starting point is poorly chosen, then, even when the optimizing process is carried out properly, it may result in a poor model, because it finds only a local optimum, which in global terms may be very poor. This raises the question of how such an important role could be played by *a priori* knowledge. In the following I will identify the relevant *a priori* knowledge as molecular structure and dynamics, and explain why 'a priori' is not as odd a description as it may sound. The second feature is the reference to Popper's falsificationist methodology. Now Popper's falsificationism is also canvassed by Barry Carpenter as the underlying methodology by which reaction mechanisms are tested by kinetic data (Carpenter 1984, Chapter 1), but falsificationism is often criticised for being too negative: in particular, Popper's scepticism about induction precludes experiments providing positive support for any kind of generalisation, and therefore knowledge of the future behaviour of any experimental system. For that very reason, it is often regarded as failing to explain how experimental knowledge can be applied in practical contexts, including engineering (see for instance Putnam 1974, Section 2). Michael Weisberg instead identifies eliminative induction as the framework for understanding how reaction mechanisms are confirmed (see Weisberg, Needham and Hendry 2011, Section 5.2.). This is a choice with which I would agree, if one must pick one of the classical conceptions of scientific method. But the main difference between eliminative induction and Popper's falsificationism is that the former, but not the latter, requires that a small handful of possible alternative theories be identified at the start of the testing process, all but one of these theories are then eliminated by experiment. How is that small handful identified? That is a crucial question, because the efficacy of eliminative induction is very sensitive to the choice. In the

next section I will argue that this role too is played by molecular structure and dynamics.

### 3 Kinetics and Reaction Mechanism

What is a kinetic model of a chemical reaction? It might be taken simply to be a mathematical expression of how the amounts (or concentrations) of the various reactants and products vary over time. Consider the following reaction, in which substances A and B react in the proportions  $w:x$  to form the products C and D in the proportions  $y:z$  (these proportions are the *stoichiometry* of the reaction):



The rate of the reaction is just the rate at which A and B are used up, or the rate at which C and D are generated. Since the reaction cannot proceed if there is no A or B present, the reaction rate must depend in some way on the amounts of A and B present (or rather, on their concentrations  $[A]$  and  $[B]$ ), but the actual dependence is expressed in the *rate law*:

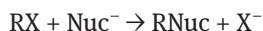
$$\text{Rate of reaction} \propto [A]^n[B]^m$$

In a simple world, the dependence of rate on reactant concentrations would reflect just the stoichiometry of the reaction:  $n$  and  $m$  would just be  $w$  and  $x$ . But as kinetics textbooks always point out, this is only rarely the case, and the order of the reaction (i.e.  $n$  and  $m$ ) must be identified experimentally (see for instance Sykes 1981, 39–40). So is a kinetic model just an empirically-based rate law? And is the rate law just an expression of the above form with the parameters  $n$  and  $m$  filled in by inspecting the *actual* time dependence of the concentrations of the relevant species? No: firstly, this would be of little use, as it provides no information about how the system would evolve under slightly different conditions. That requires some knowledge of the *dependence* of the rate of reaction on various determining factors, whether physical (e.g. temperature and pressure) or chemical (the amounts or concentrations of the reactants). Secondly, under certain conditions a reaction may display a *pseudo-order* dependence. That is, the rate of reaction may appear to depend on (or be independent of) the concentration of one of the reactants in a way that it would not under normal conditions. Thus, for instance, a reaction might ‘really’ be first order in A (i.e. the rate proportional to  $[A]$ ), but if A is in vast excess over the other reactants its concentration will

remain relatively constant and the dependence would be masked. I take it that the distinction between order and pseudo-order would not make sense if the kinetic model were just the actual variation of the concentrations of the reactants and products over time.

In fact the understanding of reaction kinetics, and the distinction between order and pseudo-order, is tied intimately to knowledge of reaction mechanism. So what is a reaction mechanism? William Goodwin identifies two conceptions at work in chemical explanation. On the *thick* conception, a reaction mechanism is ‘roughly, a complete characterization of the dynamic process of transforming a set of reactant molecules into a set of product molecules’ (2011, 310). This would involve something like a ‘motion picture’ that ‘traces, as a continuous path, the motions of the atomic nuclei’ (2011, 310). On the *thin* conception, mechanisms are ‘discrete characterizations of a transformation as a sequence of steps’ (2011, 310). The steps in question fall into a relatively small number of basic kinds: an atom or group of atoms leaving a molecule, or joining a molecule. It is the thin conception that underwrites kinetic explanation: a reaction can only proceed as fast as its slowest step – the rate-determining step – and the rate will tend to depend only on the concentrations of species involved in this step.

Consider a textbook example: the reaction of an alkyl halide  $RX$  (for instance bromoethane) with a nucleophilic ion  $Nuc^-$  (for instance the hydroxyl ion  $OH^-$ ):



Depending on the nature of the alkyl group  $R$ , the ‘leaving group’  $X^-$ , and the conditions under which the reaction takes place (e.g. the nature of the solvent), the reaction may proceed via two different mechanisms, resulting in two different rate laws. In the  $S_N1$  mechanism ( $‘S_N1’$  meaning ‘unimolecular nucleophilic substitution’),  $RX$  first dissociates (slowly) into  $R^+$  and  $X^-$  (this is the unimolecular rate-determining step), and then combines (quickly) with the nucleophile. The reaction can be expected to be first order in  $RX$ , and the rate independent of nucleophile concentration (i.e. zero order). The bimolecular  $S_N2$  mechanism, in contrast, proceeds via a mechanism in which, in a single concerted step, the nucleophile attacks the alkyl group and the leaving group departs. Since the reaction requires a molecular collision between the nucleophile and the alkyl halide, the rate can be expected to be proportional to both  $[RX]$  and  $[Nuc^-]$ .

How does this relate to the earlier theme of molecular structure and dynamics providing a framework that delimits the possibilities at the molecular level? It is simply that, for a given reaction, there will be a limited number of structurally possible ways to get from the reactants to the products. Using detailed knowledge of the structure and dynamics of molecules it will be feasible to work out these

possible mechanisms and conduct a series of experiments to determine which mechanism is realised under which conditions. As Sykes puts it, '[N]o reaction mechanism can ever be *proved* to be correct!' but

Sufficient data can nevertheless usually be gathered just to show that one or more theoretically possible mechanisms are just not compatible with the experimental results, and/or to demonstrate that of several alternatives one is a good deal more likely than the others. (1981, 43)

Roald Hoffmann gives a beautiful illustration of this process (1995, Chapter 29), describing three possible mechanisms for the photolysis of ethane to ethylene, and how H. Okabe and J. R. McNesby used isotopic labelling to eliminate two of them. By studying the kinetics of a reaction and the structure of its products (and any intermediates), and using further techniques like isotopic labelling, it is often possible to conduct a series of experiments which are collectively *crucial*, in the traditional philosophical sense that they pick out one of the various theoretical possibilities as the actual. This brings us back to eliminative induction, and the question of how *a priori* knowledge can delimit the structural possibilities in this way. The relevant body of structural theory has its origin in the 1860s, and has always been expressed through the medium of visual images rather than mathematical equations (Rocke 2010). It is perhaps one of the securest and longest-lived bodies of knowledge in science. According to G. N. Lewis, 'No generalization of science, even if we include those capable of exact mathematical statement, has ever achieved a greater success in assembling in simple form a multitude of heterogeneous observations than this group of ideas which we call structural theory' (Lewis 1923, 20–21). As a body of explanatory theory it is, of course, under empirical control, but one of its great explanatory advantages has always been that it is also under the control of *a priori* spatial intuition. It is spatial intuition that tells us that, in identifying a range of alternative mechanisms, we have exhausted the possibilities. It is only given this (fallible) judgement that we can have good reason to believe that, among the possible mechanisms we have thought of, the one best supported by the evidence (or least undermined by it) has any real chance of being correct. It is more than just the last one standing: the best of a bad bunch.



## References

- Brakel, J. van (2000). *Philosophy of Chemistry: Between the Scientific and the Manifest Image*. Leuven: Leuven University Press.
- Brakel, J. van (2011). Chemical engineering science. In: Hendry, R. F., Needham, P., & Woody, A. I. (eds.). *Handbook of the Philosophy of Science Volume 6: Philosophy of Chemistry*. Amsterdam: Elsevier. 533–547.
- Buchwald, J. (1994). *The Creation of Scientific Effects: Heinrich Hertz and Electric Waves*. Chicago: University of Chicago Press.
- Carpenter, B. (1984). *Determination of Organic Reaction Mechanisms*. New York: Wiley.
- Goodwin, W. (2011). Mechanism and chemical reaction. In: Hendry R. F., Needham, P., & Woody, A. I. (eds.) *Handbook of the Philosophy of Science 6: Philosophy of Chemistry*. Amsterdam: Elsevier. 309–327.
- Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.
- Hoffmann, R. (1995). *The Same and Not the Same*. New York: Columbia University Press.
- Lewis, G. N. (1923). *Valence and the Structure of Atoms and Molecules*. Washington, DC: Chemical Catalogue Company.
- Marquardt, W. (2013). Identification of Kinetic Models by Incremental Refinement. This volume.
- Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.
- Putnam, H. (1974). The “corroboration” of theories. In: Schilpp, P. A. (ed.). *The Philosophy of Sir Karl Popper* Volume I. LaSalle, IL: Open Court.
- Rocke, A. (2010). *Image and Reality: Kekulé, Kopp and the Scientific Imagination*. Chicago: University of Chicago Press.
- Sykes, P. (1981). *A Guidebook to Mechanism in Organic Chemistry*. Fifth Edition. London: Longmans.
- Weisberg, M., Needham, P., & Hendry, R. F. (2011). Philosophy of Chemistry. In: Zalta, E. N. (ed.). *The Stanford Encyclopedia of Philosophy* (Summer 2011 Edition). Available at <http://plato.stanford.edu/archives/sum2011/entries/chemistry/>.

### Prof. Dr. Robin Findlay Hendry

Durham University  
 Department of Philosophy  
 50 Old Elvet  
 Durham DH1 3HN  
 United Kingdom  
[r.f.hendry@durham.ac.uk](mailto:r.f.hendry@durham.ac.uk)



Valerio Lucarini

# Modeling Complexity: The Case of Climate Science

*...Chaos is the future  
And beyond it is freedom  
Confusion is next and  
Next after that is the truth...<sup>1</sup>*

*...Illud in his quoque te rebus cognoscere avemus,  
corpora cum deorsum rectum per inane feruntur  
ponderibus propriis, incerto tempore ferme  
incertisque locis spatio depellere paulum,  
tantum quod momen mutatum dicere possis.  
Quod nisi declinare solerent, omnia deorsum,  
imbris uti guttae, caderent per inane profundum,  
nec foret offensus natus nec plaga creata  
principiis: ita nil umquam natura creasset...<sup>2</sup>*

## 1 Introduction

The climatic system is constituted by four intimately interconnected sub-systems, atmosphere, hydrosphere, cryosphere, and biosphere, which evolve under the action of macroscopic driving and modulating agents, such as solar heating, Earth's rotation and gravitation (Peixoto and Oort 1992). The climate system features many degrees of freedom – which makes it *complicated* – and nonlinear interactions taking place on a vast range of time-space scales accompanying sensitive dependence on the initial conditions – which makes it *complex*. In Table 1 we present some simple examples aimed at clarifying the difference between complex and complicated systems. The distinction between these two concepts is further clarified by considering the origin of the two words: “complex” comes from the past participle of the Latin verb *complexor*, *-ari* (to entwine), whereas “complicated” comes from the past participle of the Latin verb *complico*, *-are* (to put together).

---

<sup>1</sup> Moore, T. (1983) Confusion is next. In: Sonic Youth. *Confusion is sex*. Distributed by Neutral.

<sup>2</sup> Lucretius. *De Rerum Natura* II. 216–224.

**Table 1:** Complex vs. complicated systems. Examples from natural sciences.

	Not complex	Complex
Not complicated	Harmonic oscillator	Lorenz 63 model
Complicated	Gas of non-interacting oscillators (e.g. phonons)	Turbulent fluid

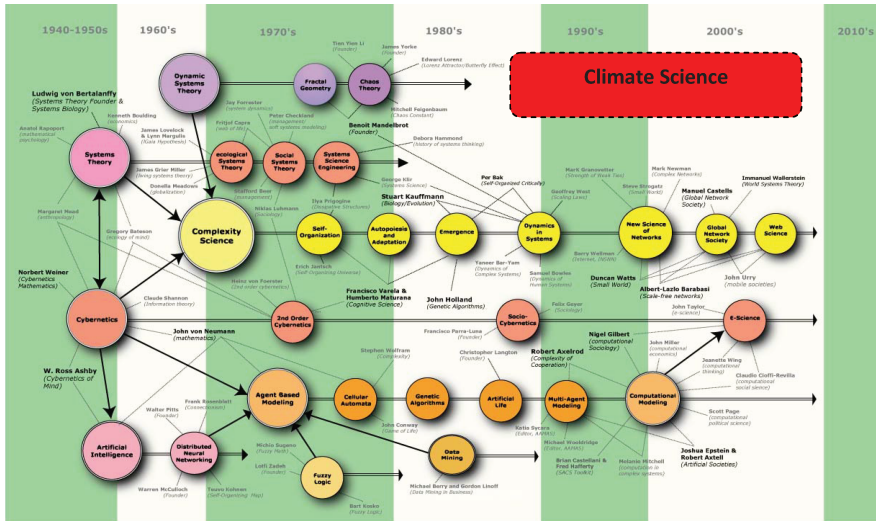
The description of the macroscopic dynamics of the climate system is based on the systematic use of dominant *balances* derived on a *phenomenological* basis in order to *specialize* the dynamical equations. Such balances are suitable classes of approximate solutions of the evolution equations representing a reasonably good approximation to the actual observed fields when sufficiently large spatial or temporal averages are considered (Speranza and Lucarini 2005). Actually, different balances have to be considered depending on the time and space scales we are focusing our interest on. Depending on the time scale of interest and on the problem under investigation, the relevant *active* degrees of freedom (mathematically corresponding to the separation between the *slow manifold* and the *fast manifold*), needing the most careful representation, change dramatically. For relatively short time scales (below 10 years) the atmospheric degrees of freedom are active while the other sub-systems can be considered essentially *frozen*. For longer time scales (100–1000 years) the ocean dominates the dynamics of climate, while for even longer time scales (over 5000 years) the continental ice sheet changes are the most relevant factors of variability (Saltzman 2002).

Such an approach reflects the fundamentally heuristic/inductive nature of the scientific research in this field, where the traditional reductionist scientific method is not necessarily effective. Climate science is a quickly evolving subject resulting from the intersection of a growing number of disciplines, such as:

- Meteorology, Oceanography, Remote Sensing, Radiative Transfer;
- Statistical Physics, Thermodynamics, Fluid Dynamics;
- Chaotic and Stochastic Dynamical Systems;
- Statistics, Data Assimilation, Data reconstruction from Proxy indicators;
- Numerical Methods, Modeling, Coding;
- Biology, Ecology, Geochemistry.

In recent years, several authors have attempted the systematization of the growing body of research dealing with complex systems under the label of *Complexity*. Numerous books and journals are being published under this, rather successful, brand, and it is encouraging to see an ever increasing degree of collaboration and exchange between social and natural scientists. For an outstanding example of such integrated activities, we refer to the FuturICT EU flagship proposal (<http://>

www.futurict.eu/). In Fig. 1 we report an example of a map of complexity, which tries to underline that the degree of interconnection between different subfields is such that the unique scientific framework of complexity can be defined. Interestingly, most if not all of the proposed maps of complexity feature a notable absence, more precisely that of climate science.



**Figure 1:** Adapted from the Castellano's complexity map (<http://www.personal.kent.edu/~bcastel3/>). The balloon indicating Climate Science has been added by the author.

One could propose that in the map presented in Fig. 1, the large empty space in the upper right corner should be filled with a balloon referring to climate science. Such an absence is considerably puzzling if one considers that some of the most notable features shared by most complex systems (e.g. sensitive dependence on initial conditions, multiscale properties, intermittency) have been discovered in the context of or in vicinity to climate problems, and that climate science, especially in the last two decades, has emerged to being one of the most widely discussed scientific fields. Maybe this is actually the reason of such a notable absence: climate science is perceived as a mostly policy-driven, high-tech, computer muscled-up field, rather than being a frontier subject where to test and improve the refined tools and concepts needed to analyze and deal with complexity.

Therefore, it is clear that the investigation of the global structural properties plays a central role for the provision of a unifying picture of the climate system.

Such an endeavor is of fundamental importance for improving substantially our understanding of climate variability and climate change on a large variety of scales, which encompass major paleoclimatic shifts, almost regularly repeated events such as ice ages, as well as the ongoing and future anthropogenic climate change, as envisioned by the scientific program proposed in the landmark book by Saltzman (2002).

Such an effort has significant relevance also in the context of the ever-increasing attention paid by the scientific community to the quest for validating climate models (CMs) of various degrees of complexity, as explicitly requested by the Intergovernmental Panel on Climate Change (IPCC) in its 4<sup>th</sup> Assessment Report (IPCC 2007), and for the definition of strategies aimed at the radical improvement of their performance (Held 2005; Lucarini 2008a).

In a modern, global perspective, the climate can be seen as a complex, non-equilibrium system, which generates entropy by irreversible processes, transforms moist static energy into mechanical energy as if it were a heat engine, and, when the external and internal parameters have fixed values, achieves a steady state by balancing the input and output of energy and entropy with the surrounding environment (Peixoto and Oort 1992, Johnson 2000, Lorenz and Kleidon 2005, Lucarini 2009a). The tools of phenomenological non-equilibrium thermodynamics (de Groot and Mazur 1962) seem very well suited in defining a new point of view for the analysis of the CS for understanding its variability and its large-scale processes, including the atmosphere-ocean coupling, the hydrological cycle, as well as understanding the mechanisms involved in *climate phase transitions* observed at the so-called tipping points (Lenton et al. 2008), i.e. conditions under which catastrophes may occur for small variations in the boundary conditions or in the internal parameters of the system (Fraedrich 1979).

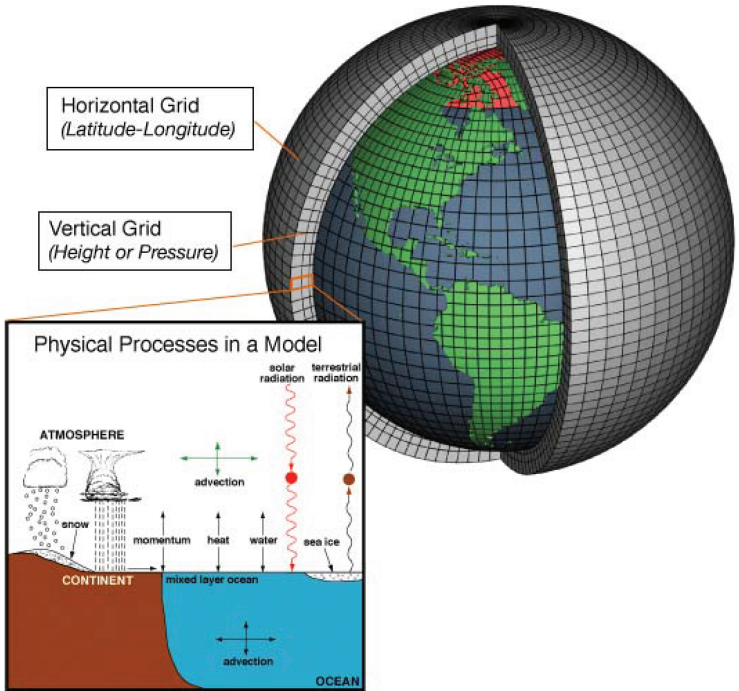
Moreover, a primary goal of climate science is to understand how the statistical properties – mean values, fluctuations, and higher order moments – of the climate system change as a result of modulations to the parameters of the system occurring on various time scales. A large class of problems fall into this category, such as those involving climate sensitivity, climate variability, climate change, climate tipping points, as well as the response to daily, seasonal, orbital forcings, to changes in the atmospheric composition, to changes in the geography and topography of the continents and of the seafloor. Recent results from non-equilibrium statistical mechanics mostly due to Ruelle (1998, 2009) provide rigorous tools for tackling this problem using a perturbative approach (Lucarini 2008b, 2009b; Lucarini and Sarno 2011).

## 2 Issues in Climate Modeling

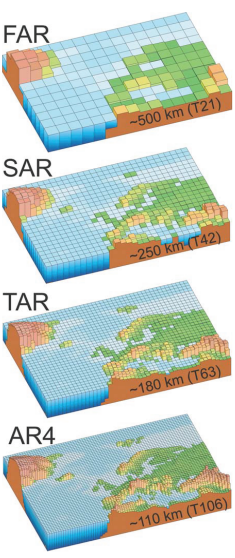
Given the nature of their research, numerical simulation has been a key method of investigation for climate scientists since the early days of computers. Actually, in the late 1940s, the first large-scale application of automatic computing consisted in the first numerical weather forecast, based on greatly simplified equations, which was proposed by Von Neumann and mainly devised by Charney. This also emphasizes the long-standing strategic relevance of climate-related science. Since the late 1950s, the US (and Swedish) technical services have been using computer-assisted numerical integration of relatively accurate equations descriptive of the physics of the atmosphere to routinely produce weather forecasts.

The evaluation of the accuracy of numerical climate models and the definition of strategies for their improvement are, today more than ever, crucial issues in the climate scientific community. On the one hand, climate models of various degrees of complexity constitute tools of fundamental importance to reconstruct and project the state of the planet in the future and to test theories related to basic geophysical fluid dynamical properties of the atmosphere and of the ocean as well as of the physical and chemical feedbacks within the various subdomains and between them. On the other hand, the outputs of climate models, and especially future climate projections, are gaining an ever-increasing relevance in several fields, such as ecology, economics, engineering, energy, and architecture, as well as for the process of policy-making at national and international level. Regarding influences at societal level of climate-related finding, the impacts of the 4<sup>th</sup> Assessment Report of the IPCC (2007) are unprecedented, to the point that the Panel was awarded the 2007 Nobel Prize for Peace.

Numerical modeling options strongly rely on the available computer power, so that the continuous improvements in both software and hardware have permitted a large increase in the performances of the models and at the same time an impressive widening of their horizons. See Fig. 2 for an overview of the structure of a state-of-the-art CM. We remind that parametrizations are approximate representation of the effects on the scales resolved by CMs of the processes occurring in the range of unresolved scales. See a modern perspective of this problem in Palmer and Williams (2010). On the one hand, the adoption of finer and finer resolutions has allowed a more detailed description of the large scale features of the dynamics, and, more critically, a more direct physical description of a larger set of processes, thus limiting the need for parameterization procedures, which, where needed, have become more accurate. On the other hand, it has been possible to implement and then refine the coupling between models pertaining to different systems having a common boundary, such as the atmosphere and the ocean, or the atmosphere and the land surface. See a pictorial representation of



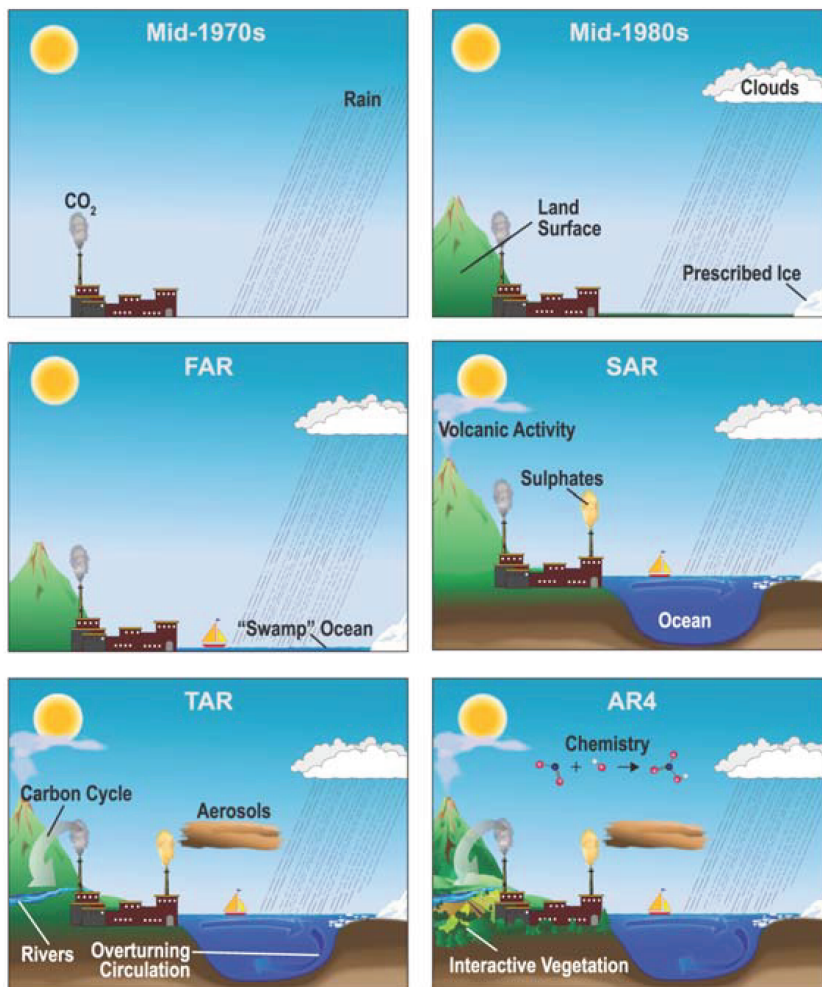
**Figure 2:** Overview of the structure of a state-of-the-art climate model. From the NOAA research website [http://www.research.noaa.gov/climate/t\\_modeling.html](http://www.research.noaa.gov/climate/t_modeling.html)



**Figure 3:** Geographic resolution characteristic of the generations of climate models used in the IPCC first (FAR, 1990), second (SAR, 1996), third (TAR, 2001), and fourth (AR4, 2007) assessment reports. These illustrations are representative of the most detailed horizontal resolution used for short-term climate simulations. Vertical resolution in both atmosphere and ocean models is not shown, but it has increased comparably with the horizontal resolution, beginning typically with a single-layer slab ocean and ten atmospheric layers in the FAR and progressing to about thirty levels in both atmosphere and ocean. From IPCC (2007), p. 113.



the improvement in the horizontal resolution of CMs between the 1980's and the 2000's in Fig. 3. The increase in the number of natural processes represented in CMs in the same time frame is represented in Fig. 4.



**Figure 4:** The number of natural processes explicitly represented in climate models has increased over the last few decades. The additional physics incorporated in the models are shown pictorially by the different features of the modeled world. From IPCC (2007), p. 99.

Still, since the climate is a multiscale system (Schertzer and Lovejoy 2004), our ability to represent it with numerical models is intrinsically limited. One should consider that climate variability is observed within spatial scales ranging from  $10^{-6}$  m to  $10^7$  m and temporal scales ranging from  $10^{-6}$  s to  $10^{16}$  s. These ranges dwarf what covered explicitly by present top-notch models by many orders of magnitude. The progress in terms of computing power of a factor of  $10^6$  obtained in the last 30 years has reduced only by a relatively small amount the distance between model and the actual system, so that it seems unfeasible to expect within the next decades fundamental progresses to our understanding of the climate system obtained only through brute force computing. This is in stark contrast to what envisioned by Navarra et al. (2010).

Climate modeling faces uncertainties belonging to two distinct classes. The uncertainties on the initial conditions (*uncertainties of the first kind*) limit, because of the chaotic nature of the system, our ability to predict deterministically the state of the system at a future time, given our imperfect knowledge of its state at the present time. In the growing body of research dealing with climate prediction, this kind of uncertainty is partially taken care of by applying the same strategies today commonly adopted in the usual weather forecasting, i.e. by using ensemble simulations. Along these lines, many simulations are started with slightly perturbed initial conditions, and the set of evolved trajectories is used to provide a probabilistic estimate of how the system will actually evolve. Obvious limitations are related to the technological difficulties of running a sufficient number of ensemble members. But more basic problems are also present. The structural deficiencies together with an unavoidably limited knowledge of the external forcings (*uncertainties of the second kind*) limit intrinsically the possibility of providing realistic simulations of the statistical properties of the climate system, especially affecting the possibility of representing abrupt climate change processes.

The validation, or auditing – overall evaluation of accuracy – of a set of climate models, is a delicate operation, which can be decomposed in two related, albeit distinct, procedures. The first procedure is the intercomparison, which aims at assessing the consistency of the models in the simulation of certain physical phenomena over a certain time frame. The second procedure is the verification, whose goal is to compare the models' outputs to corresponding observed or reconstructed quantities. Hence, a third kind of uncertainty is related to the actual procedure of auditing: what are the best *metrics*, i.e. the best statistical estimators to be used for analyzing the output of climate models? In principle, any reasonable function of the variables included in our climate model is a perfectly legitimate metrics. Nevertheless, even if all such *observables* are mathematically well defined, their physical relevance and robustness can be very different. Since

no strict a-priori criterion exists for selecting a good observable, even if taking into account some basic physical properties of the climate system can provide useful guidance, as explained below, we do not have a unique recipe for testing our models. Again, this is in stark contrast to the case of more traditional scientific fields, where the relevant observables (e.g., in high-energy physics, “mass”, “transition probability”, “cross-section”) are suggested by the very equations we are trying to solve or analyze experimentally.

### 3 Performance Metrics and Uncertainties

A matter of great interest in the analysis of climate models is the choice of the physical observables used in the auditing procedures, or, as they are often referred to, of the metrics of validation of the climate models. An ever-increasing attention is being paid by the scientific community to the quest for reliable, robust metrics, as explicitly requested by the 4<sup>th</sup> Assessment Report of the IPCC.

Most typically, the models’ validation is based upon the analysis of the skill in simulating fields of common practical interest, such as the surface air temperature or the accumulated precipitation. However, these fields describe quantities that can hardly be considered *climate state variables*. By considering the vertical profile of the annual and global mean temperature, the zonal mean surface air temperature, or precipitation, the impression is that all models have very similar performances and it is very difficult to assess whether a model is performing in any sense better than any other. Nevertheless, they differ substantially in the horizontal as well as vertical resolution, numerical schemes, physical parameterizations, and so on.

One aim – from the end-user’s point of view – is immediately checking how *realistic* the modeled fields of practical interest are. But if the aim is to define strategies aimed at the radical improvement of their performance, beyond incremental advances often obtained at the price of large increases in requested computed power, it is important to fully understand the differences in the representation of the *climatic machine* among models and possibly decide whether specific physical processes are correctly simulated by a specific model.

In order to analyze the representation of specific physical processes as well as of balances involving conservation principles, it is necessary to use specialized diagnostic tools – that we may call *process-oriented metrics* – as indexes for model reliability. Such an approach may be helpful in clarifying the distinction between the performance of the models in reproducing *diagnostic* and *prognostic* variables of the climate system. The definition of efficient process-oriented

metrics benefits from the adoption of a well-defined scientific framework. In section 6 we propose our point of view where we maintain that a thermodynamic perspective is well-suited for analyzing the climate system, because it provides a way to cut through its complexity and, at the same time, carefully takes into account its non-equilibrium properties.

Additional practical as well as epistemological issues emerge when we consider the actual process of comparing theoretical and numerical investigations with experimental data. Model results and approximate theories can often be tested only against observational data from the past, which may feature problems of various degrees of criticality, essentially because of the physical extension of the systems under analysis. The available historical observations sometimes feature a relatively low degree of reciprocal synchronic coherence and individually present problems of diachronic coherence, due to changes in the strategies of data gathering with time, whereas proxy data, by definition, provide only semi-quantitative information on the past state of the climate system. The natural variability of both the model and the real system contributes to blur the line between a failed and a passed test. Anyway, a positive result would not at all ensure the model's ability to provide consistent future projections, whereas at most it is possible to deduce from a negative result that the model is not reliable enough. Summarizing, difficulties in the process of falsification basically emerge because we always have to deal with three different kinds of attractors: the attractor of the real climate system, its reconstruction from observations, and the attractors of the climate models.

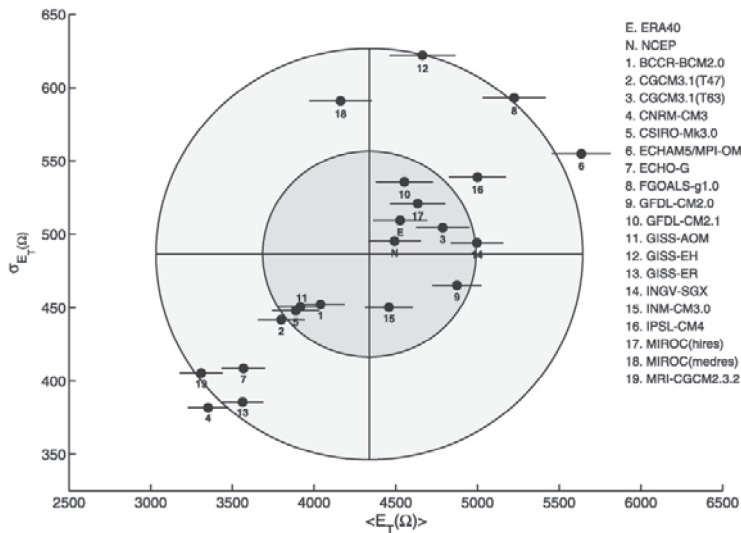
The unavoidable presence of such critical uncertainties implies that every model used to generate projections about future climate change could be interpreted as being weak in its descriptive power. When considering the climate as a whole, there are no real or virtual laboratories, where theories and models can be tested in a classic sense against experiments, because, due to the entropic time arrow, repeatability is strictly not possible. Using a standard scientific procedure would imply to reject a model, if it fails to comply with even just one observable. That is how, e.g., high-energy physics typically works, as shown by the very idea of building the Large Hadron Collider (LHC). As clear from the previous discussion, it is unfeasible to use this criterion in climate science, because we would end up discarding all models and arresting any progress. Therefore, the Galilean scientific framework given by recurrent interplay of experimental results and theoretical predictions is challenged.

As for taking care of possible issues related to initial conditions, often one considers an ensemble of simulations, where the same climate model is run under identical conditions from slightly different initial state. This allows a more

detailed exploration of the phase space of the system, with a better sampling – on a finite time – of the attractor of the model.

The deficiencies of a single climate model and the stability of its statistical properties can be addressed by applying Monte Carlo techniques to generate an ensemble of simulations, each characterized by different values of some uncertain key parameters characterizing the global climatic properties. Therefore, in this case, sampling is performed by considering attractors that are parametrically deformed, which is, by the way, a formally well-defined operation when we consider the Ruelle (1998, 2009) response theory (see discussion in Lucarini 2008b).

A detailed analysis of structural uncertainties requires the comparison of different models following a horizontal and vertical conceptual hierarchical path. The horizontal comparison is the comparative study of the results generated by models sharing a roughly common level of complexity, but having been implemented in different ways by different people. The vertical comparison is the comparative study of the results obtained by a family of models, each built as an extension and complexification of another one starting from an initial simple parent, thus creating a natural hierarchy of increasing complexity.



**Figure 5:** Performance of various state-of-the-art climate models in representing winter mid-latitude northern hemisphere atmospheric variability. Climatology of integrated spectral power of the waves is plotted against its interannual variability (in  $\text{m}^2\text{s}^{-2}$ ). The various points correspond to the data set indicated in the legend. The ensemble mean is located at the centre of the ellipses. Adapted from Lucarini et al. (2007).

The Project for Climate Model Diagnostics and Intercomparison (PCMDI), through its climate models intercomparison projects (CMIPs), has supported the gathering into a single web-location of climate model outputs contributing to the activities initiated by the IPCC. The PCMDI thus provides a unique opportunity for evaluating the state-of-the-art capabilities in simulating the behavior of climate system. The CMIP has provided a rather complete and standardized set of climate outputs in its third phase, which was related to the IPCC (2007) report, whereas the CMIP's fifth phase will collect data relevant for the preparation of the fifth assessment report of IPCC.

In order to describe synthetically and comprehensively the outputs of a growing number of climate models, recently it has become common to consider multi-model ensembles and focus the attention on the ensemble mean and the ensemble spread of the models, taken respectively as the (possibly weighted) first two moments of the models outputs for the considered metric. Then, information from rather different attractors is merged. Whereas this procedure surely has advantages, such statistical estimators should not be interpreted in the standard way – the mean approximating the truth, the standard deviation describing the uncertainty – because such a straightforward perspective relies on the (false) assumption that the set is a probabilistic ensemble, formed by equivalent realizations of a given process, and that the underlying probability distribution is unimodal. Figure 5 portrays the statistical properties of the mean value (x-axis) and interannual variability (y-axis) of a quadratic measure of the strength of winter northern hemisphere mid-latitude atmospheric disturbances during the period 1961–2000 and reports the results for 19 state-of-the-art climate models included in the PCMDI dataset. Moreover, reference data are reported for the two reanalyses datasets, produced by NCEP-NCAR and ECMWF, commonly considered as roughly equivalent reconstructions of the true atmospheric state. As we see, the ensemble mean (centre of the two ellipses) is actually rather close to the “true” state, but, on the other hand, it is positioned in a location where the density of the points referring to the outputs of the various models is very low. Note that the two semi-axes of the internal (external) ellipsis are given by (twice) the values of the standard deviation of the ensemble for the two considered variables. Therefore, it is at least questionable to interpret the ensemble mean as representative in any well-defined sense of the models' outputs.

## 4 A Side Note: An Outlook on Geo-Engineering

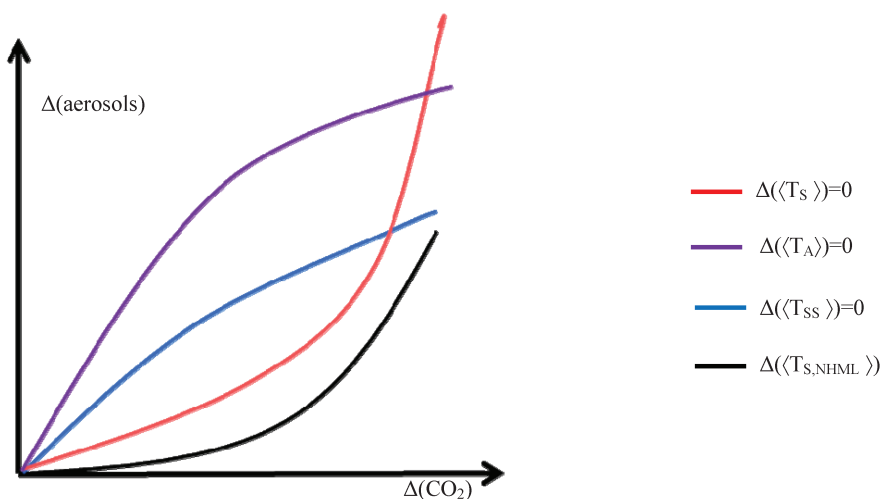
In spite of all the efforts of several scientific communities, pressure groups, and citizens, the more and more widespread concern regarding the climatic impacts of the observed steady increase of CO<sub>2</sub> concentration in the atmosphere has not been met with the actual provision of an effective, international and multilateral protocol of economic and political measures aimed at limiting present and future hazards. Being able to deal effectively, and in the context of an increasingly multipolar world, with the complexities of the global economy and politics and of the global climate system at the same time seems an almost insurmountable task. In this context, in recent years a growing number of scientists, policy-makers, and corporations have proposed the adoption of geo-engineering strategies as – at least – short-term mitigation of climate change effects due to CO<sub>2</sub> increase.

In general, geo-engineering refers to the adoption of measures aimed at modifying, on purpose, the climate system in a – allegedly – controlled way. On smaller temporal and spatial scales, several weather modification strategies have been devised in the course of the years, such as the seeding of clouds aimed at increasing their rain efficiency. Nevertheless, geo-engineering is distinct as its scope is intrinsically global in space and multiannual in time. One of the most relevant proposals in this direction has been that of continuously injecting in the atmosphere large amounts of aerosols in order to reduce the amount of net incoming solar radiation (some aerosols reflect the solar radiation quite efficiently), thus countering the anthropogenic greenhouse effect due to ever-increasing CO<sub>2</sub> concentration. This idea has been evaluated as technologically feasible and economically very convenient with respect to challenging the present model of economic development. Putting aside the ethical issues related to the idea of countering pollution with further pollution and those to the fact that a single country or, in principle, even a private can decide to alter unilaterally the global climate, and neglecting the large scientific uncertainties still surrounding the actual effects of such large-scale injection of material in the atmosphere, the complexity of the climate system seems to suggest that this kind of operation is intrinsically ill-posed, or better, far from being able to provide a simple solution to a complex problem like global warming.

Mathematically, we can say that geo-engineering is about defining suitable isolines constructed in the following way: If we consider an increment  $x$  of CO<sub>2</sub> concentration, what is the amount of aerosols  $y$  needed to keep the average value of the statistical properties of the climate variable  $z$  constant? By changing the value of  $x$  and finding the corresponding values of  $y=y_z(x)$ , we construct the isoline of the climate variable  $z$ , i.e., when moving parametrically along such a line (corresponding to the adoption of geo-engineering measures contrasting



the increase in  $\text{CO}_2$  concentration), the climatology of  $z$  is not altered. But, if we choose any other climate variable  $z_1, z_2, \dots, z_n$ , the geo-engineering strategy will not provide any solution, because  $z_1, z_2, \dots, z_n$  are, instead, constant along the isolines  $y=y_{z_1}(x), \dots, y=y_{z_n}(x)$ , which are in general distinct from  $y=y_z(x)$ . Therefore, along the parametric curve  $y=y_z(x)$  the value of the climate variables  $z_1, z_2, \dots, z_n$  will definitely change, so that *climate will change*. Injecting the aerosols in the atmosphere has the effect of modulating, but not of erasing in any real sense, the effect of increasing  $\text{CO}_2$  concentrations.



**Figure 6:** Apart from the scientific uncertainties, geo-engineering measures provide a fix only for a selected climate observable. The isolines of  $\langle T_s \rangle$ ,  $\langle T_A \rangle$ ,  $\langle T_{ss} \rangle$ , and  $\langle T_{s,NHML} \rangle$  are, in fact, different. See details in the text.

Therefore, the geo-engineering strategy described by  $y=y_z(x)$  will only provide an example of a constrained climate change scenario, and not at all a scenario foreseeing the cancellation of climate change in general. We then understand that all the emphasis is in the selection of the  $z$ -variable of interest, and it seems rather clear that such a choice has an eminently political nature, and, furthermore, it seems hopeless to reach a global consensus on the “right” variable to consider in a hypothetically pro-geo-engineering world. In Fig. 6 we provide a graphical representation of this issue, where we consider four variables (globally averaged surface temperature ( $\langle T_s \rangle$ ), averaged sea surface temperature ( $\langle T_{ss} \rangle$ ), averaged atmospheric temperature ( $\langle T_A \rangle$ ) and surface temperature averaged over the land



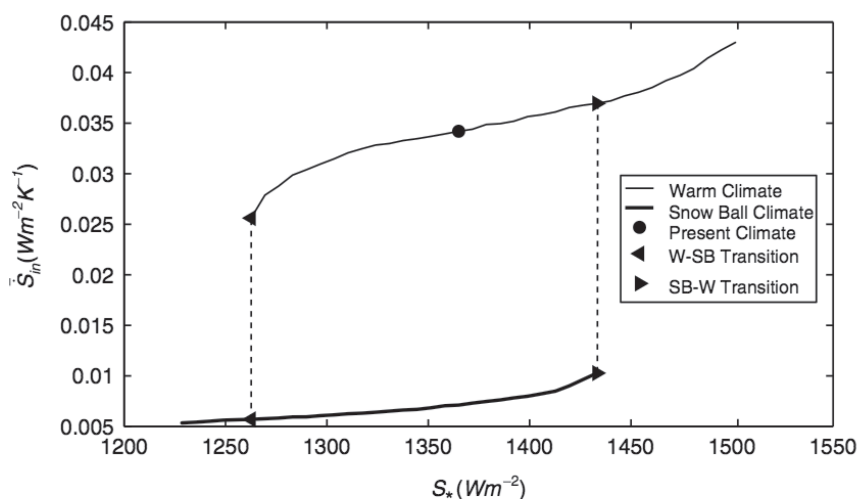
located in the mid-latitudes of the northern hemisphere ( $\langle T_{S,NHML} \rangle$ ). Each country or group of countries will have different and even contrasting interests, as the effects of climate change are felt locally and the adaptive capacity is widely different. Note that, by definition, potential strategies aimed at achieving a reduction of the  $\text{CO}_2$  concentration in the atmosphere do not suffer from this problem. Therefore, geo-engineering seems to be a logical loophole: rather than providing a practical solution to the ongoing anthropogenic forcing, it moves the difficulties to the choice of the optimization strategy.

## 5 Our Proposal: A Thermodynamic Perspective

Many authors have approached the problem of understanding the properties of the CS by studying the structure of the bifurcations of dynamical systems constructed heuristically and featuring a minimal number of climatically relevant variables (usually below 10). This strategy has led to great scientific results and suggested the existence of generic mathematical structures, sometimes re-discovered in hierarchies of CMs. A relevant example of investigation performed along these lines on processes occurring on multi-decadal time scale is the analysis of the stability of the thermohaline circulation. On atmospheric time scales, some of the most important investigations of the low-frequency variability of the mid-latitude atmosphere have been carried out along similar lines. The limitations of this approach lie in the fact that the simplifications adopted in the derivation of the dynamical systems may blur out the involved physical processes and hardly allow for an efficient representation of the fluctuations of the system, to which the introduction of stochastic forcing provides a partial solution (Hasselmann 1976). This approach suffers from need for a – usually beyond reach – closure theory for the noise properties.

While acknowledging the scientific achievements obtained along the above mentioned line, we propose a different approach for addressing the *big picture* of a complex system like climate. An alternative way for providing a new, satisfactory theory of climate dynamics able to tackle simultaneously balances of physical quantities and dynamical instabilities is to adopt a thermodynamic perspective, along the lines proposed by Lorenz (1967). We consider simultaneously two closely related approaches, a phenomenological outlook based on the macroscopic theory of non-equilibrium thermodynamics (see e.g., de Groot and Mazur 1962), and, a more fundamental outlook, based on the paradigm of ergodic theory (Eckmann and Ruelle 1985) and more recent developments of the non-equilibrium statistical mechanics (Ruelle 1998, 2009).

The concept of the energy cycle of the atmosphere introduced by Lorenz (1967) allowed for defining an effective climate machine such that the atmospheric and oceanic motions simultaneously result from the mechanical work (then dissipated in a turbulent cascade) produced by the engine, and re-equilibrate the energy balance of the climate system. One of the fundamental reasons why a comprehensive understanding of climate dynamics is hard to achieve lies in the presence of such a nonlinear closure. Recently, Johnson (2000) introduced a Carnot engine-equivalent picture of the climate system by defining effective warm and cold reservoirs and their temperatures, and deriving a suitably defined efficiency. The interest towards studying the climate irreversibility largely stemmed from the proposal of the maximum entropy production principle (MEPP), which suggests that non-equilibrium nonlinear systems adjust in order to maximize the entropy production (Ozawa et al 2003, Kleidon and Lorenz 2005). Even if recent claims of *ab initio* derivation of MEPP have been dismissed, it has stimulated the re-examination of entropy production in the climate system (Pascale et al. 2009) and the development of new strategies for improving the CMs parameterization.



**Figure 7:** Dependence of the Entropy Production of the climate system on the value of the solar constant. Note the presence of a wide region of bistability, where both the warm (W) and the snowball (SB) climates are stable. Adapted from Lucarini et al. (2010a).

Recently, a link has been proposed between the Carnot efficiency, the intensity of the Lorenz energy cycle, the entropy production and the degree of irreversibility of the climate system (Lucarini 2009a). In particular, it has been found

that the efficiency of the equivalent thermal machine sets also the proportionality between the internal entropy fluctuation of the system and the lower bound to entropy production by the fluid compatible with the 2<sup>nd</sup> law of thermodynamics. Such a bound is basically given by the entropy produced by the dissipation of the mechanical energy, whereas the excess of entropy production is due to the transport of heat down the gradient of the temperature field. These results pave the way for a new, extensive exploration aimed at understanding the climate response under various scenarios of forcing, of atmospheric composition, and of boundary conditions. Recent preliminary efforts have focused on the impacts on the thermodynamics of the climate system of changes in the solar constant, with the analysis of the onset and decay of snowball Earth conditions (Lucarini et al. 2010a), and on those due to changing CO<sub>2</sub> concentration (Lucarini et al. 2010b). In the snowball Earth experiment, the two climate regimes (ice-covered and today-like) feature radically different physical properties. In particular, the climate efficiency decreases (increases) with increasing solar constant in present (snowball) climate conditions. Moreover, entropy production (see Fig. 7) and the irreversibility of the system are much higher in warmer climates. When considering CO<sub>2</sub> changes, a warmer CS results to be less efficient, more irreversible, and produces more entropy. While in cold climates a dominating role for the changes in the thermodynamics is played by changes in the vertical stratification of the atmosphere, in warm ones changes in latent heat fluxes are crucial.

As the results in Lucarini (2009) allow for treating the exchange of mechanical energy between atmosphere and ocean as a boundary term in the energy budget, this approach may contribute to quantifying the mechanisms involved in the mechanical energy budget in the global ocean, which have long been a source of debate in oceanography (e.g. Wunsch and Ferrari 2004, Tailleux 2010). Additionally, a thermodynamic analysis of the climate transitions at the tipping points (Lenton et al. 2008) based upon macro-scale thermodynamic properties is also proposed. In Lucarini et al. (2010a) it is shown that the loss of stability of a climate regime is accompanied by the transition to a regime featuring a less efficient climate, which is characterized by thermodynamic conditions closer to equilibrium. It seems very relevant to tackle the analysis of the suggestive hypothesis of the generality of this behavior. This has implications for the issue of multiple stability in the atmosphere-biosphere system.

Recently, it has been shown that it is possible to compute the entropy production and derive information on the Lorenz energy cycle by only looking at the 2D fields of top-of-the-atmosphere and surface radiative budgets (Lucarini et al. 2011). This paves the way for studying the thermodynamics of the *climates* of planetary bodies other than the Earth, whose investigation has been, by the way, one of the first applications of MEPP (Kleidon and Lorenz 2005). This is a

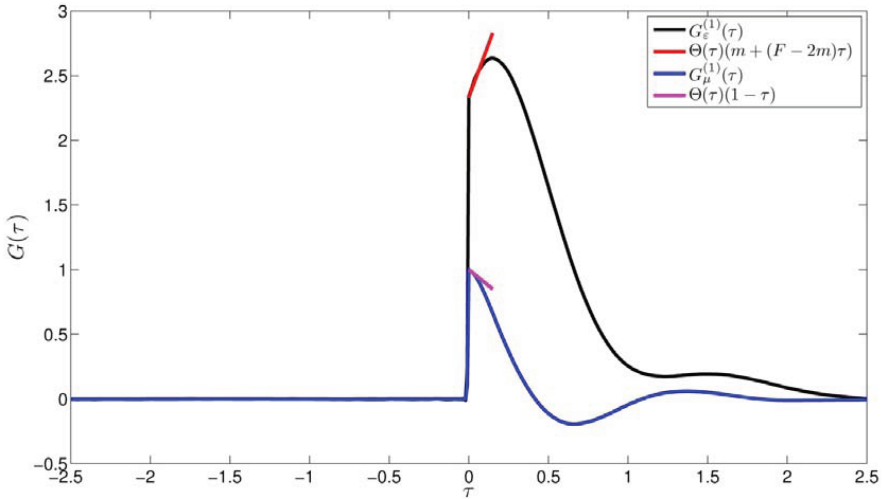
rather promising perspective, given the ever increasing attention paid to, and data obtained on, these astronomical objects. See the recent study in this direction performed by Boschi et al. (2013).

The fundamental approach based upon non-equilibrium statistical mechanics provides great opportunities, due to the recent development of the discipline (Gallavotti 2006), but also great challenges. A serious difficulty in the analysis of the CS is that the fluctuation-dissipation theorem (see, e.g., Kubo 1966), cornerstone of quasi-equilibrium statistical mechanics, cannot straightforwardly be applied, because the climate is a non-equilibrium, forced and dissipative system, where the asymptotic dynamics take place in a strange attractor. Natural fluctuations and forced motions cannot be equivalent, because – while natural fluctuations of the system are restricted to the attractor, due to the fact that asymptotically there is no dynamics along the stable manifold – external forcings will induce motions out of the attractor with probability 1 (Ruelle 1998, 2009, Lucarini 2008, Lucarini and Sarno 2011). In a climatic context, this corresponds to an earlier intuition by Lorenz (1979) on the non equivalence between forced and free fluctuations. Therefore, it is at least questionable to take for granted that climate change signals should project on the natural modes of variability.

Recently, Ruelle (1998, 2009) introduced a mathematical theory for computing *ab initio* the response of a large class of non-equilibrium systems to external perturbations. The theory specifically applies only to a specific class (Axiom A) of statistical mechanical systems. Nevertheless, accepting the chaotic hypothesis (Gallavotti 2006), this class provides an excellent model for general physical systems. More recently, it has been proved that Kramers-Kronig (KK) relations connect the real and imaginary part of the susceptibilities at all orders of nonlinearity. The Ruelle response theory provides a rigorous way to compute explicitly, as well-defined perturbation series, the climate response of a system to forcings featuring generic time modulation and generic spatial pattern. The KK theory and the related sum rules (Lucarini 2008b) can be used to define a comprehensive self-consistent theory of climate change against forcings of all time scales and constitute a formidable tool for assessing the consistency of a CMs, since they provide explicit and computable constraints, based only upon the principle of causality, that have to be necessarily obeyed. Models not complying with these constraints cannot feature a consistent dynamics over all of the time and space scales and require a detailed re-examination (Lucarini 2008b).

The analysis of these properties with CMs of various degrees of complexity seems absolutely relevant. The prototypical numerical study by Lucarini (2009b) has been extended by Lucarini and Sarno (2011), where the first direct computation of the Green function of a simplified climate model has been performed (see Fig. 8). The Green function allows for computing *ab initio* the response of the

considered climate observable to the external perturbation introduced into the system. This provides a promising way to compute probabilistic climate projections. Further theoretical extensions and applications to models of higher complexity and deeper climatic interest are definitely necessary.



**Figure 8:** Green function describing the response to a specific perturbation of the spatially averaged total energy (black line) and total momentum (blue line). The short-term behavior, computed *ab initio* using the response theory, is indicated in red (energy) and magenta (momentum). Adapted from Lucarini and Sarno (2011).

Using the response theory formalism and its extension to the frequency domain, it is possible to compute the climatic impact of quasi-static perturbations, such as those related to changes in the parameters of the system, like atmospheric composition, albedo, solar irradiation or Earth's axis inclination. Moreover, it is possible to tackle rigorously issues such as determining the impact of periodic forcings like the seasonal cycle, the solar cycle, and multi-millennial orbital variations. As in quasi-geostrophic atmospheric modeling, the anomalies in topography and surface temperature appear as boundary conditions terms controlled by (small) parameters, one can compute explicitly their impact on the statistical properties of the circulation, thus extending the work of Speranza et al. (1985) on orographic modification to baroclinic instability in a climatic perspective. Similar strategy could be used for specific oceanic problems. Moreover, Wouters and Lucarini (2012) have shown how response theory provides a well-defined strategy for deriving rigorous deterministic and stochastic parameterizations for unresolved processes.

Finally, the analysis of the susceptibility function can highlight and quantify relevant climate feedbacks. In fact, the response of the system varies enormously with the time scale of the forcing: resonances with the internal time scales may greatly amplify the response to perturbations. On a similar note, the analysis of tipping points, as conditions under which the susceptibility diverges, could be envisioned.

## 6 Conclusions

We have briefly recapitulated some of the scientific challenges and epistemological issues related to climate science. We have discussed the formulation and testing of theories and numerical models, which, given the presence of unavoidable uncertainties in observational data, the non-repeatability of world-experiments, and the fact that relevant processes occur in a large variety of spatial and temporal scales, require a rather different approach than in other scientific contexts.

In particular, we have clarified the presence of two different levels of unavoidable uncertainties when dealing with climate models, related to the complexity and chaoticity of the system under investigation. The first is related to the imperfect knowledge of the initial conditions; the second is related to the imperfect representation of the processes of the system, which can be referred to as structural uncertainties of the model. We have discussed how Monte Carlo methods provide partial but very popular solutions to these problems. A third level of uncertainty is related to the need for a, definitely non-trivial, definition of the appropriate metrics in the process of validation of the climate models. We have highlighted the difference between metrics aimed at providing information of great relevance for the end-user from those more focused on the audit of the most important physical processes of the climate system.

It is becoming clearer and clearer that the current strategy of incremental improvements of climate models is failing to produce a qualitative change in our ability to describe the climate system, also because the gap between the simulation and the understanding of the climate system is widening (Held 2005, Lucarini 2008a). Therefore, the pursuit of a “quantum leap” in climate modeling – which definitely requires new scientific ideas rather than just faster supercomputers – is becoming more and more of a key issue in the climate community (Shukla et al. 2009). In this context, we could not disagree more with the perspective of climate science proposed in Navarra et al. (2010), who foresee a dominance of supercomputing in few selected centers, central planning of scientific priorities, and reorganization of whole academic and scientific framework in close resemblance

with what was done in high-energy physics over 50 years ago. First, centralized planning of the scientific priorities (with the related allocation of funds and jobs) automatically raises the question of who is going to define such priorities and on which basis. Second, and more importantly, as widely discussed in this paper, it is hard to find scientific sectors with as different epistemologies as high-energy physics and climate science. Navarra et al. (2010) talk about “crucial experiments”, but, unfortunately, these just cannot exist in a non-Galilean setting as that of climate science. In fact, the distance of climate science from the “timeless” Galilean science based upon repeated cycles of experimental investigations and improvements to scientific theory is so wide that it is impossible to apply the usual scientific validation criteria to the results of climate science. The different epistemology pertaining to climate science implies that its answers cannot be singular and deterministic, while they must be plural and stated in probabilistic terms. Flexible and open-source modeling, such as that represented by the PLASIM platform (Fraedrich et al. 2005), and distributed computing, such as that adopted in the climaprediction.net project (Allen 1999), seem in principle more suited for the goals, the methodologies, and the development of climate science. Moreover, proposing new ideas, innovative scientific frameworks, and new paradigms, rather than flexing and training metaphorical (and expensive) muscles, seems definitely more promising in the author’s view.

In this regard, we have proposed the adoption of a thermodynamic perspective as a potentially relevant framework for improving our understanding of the climate system and our ability to model it. The macroscopic non-equilibrium thermodynamics allows for characterizing the climate system in terms of its efficiency to produce work, i.e. organized atmospheric and oceanic motion, to achieve steady state by balancing the input and output of energy and entropy with the surrounding environment, and of its irreversibility, due to entropy-generating processes. Such global properties allow for diagnosing, describing, and understanding the smaller scale processes associated to climate variability, climate feedbacks, climate change in general, and large scale climate re-organizations occurring at tipping points in particular. Moreover, these tools can be used for studying the basic properties of the circulation of planetary atmospheres, a topic of great interest in the present age characterized by the discovery of quickly growing number of exoplanets.

A more fundamental approach, based upon non-equilibrium statistical mechanics, can also be envisioned. The fact that, as can be deduced from Ruelle’s (1998, 2009) arguments, the climate system does not obey the fluctuation dissipation theorem, is another crucial reason why its modeling and its understanding are intrinsically difficult. The climate responses to forcings are in principle irreducible to internal fluctuations. Therefore, as opposed to common wisdom

in climate science, it is not obvious at all that, *e.g.*, climate change signals will project on natural modes of variability. Nonetheless, one should also consider that if stochastic forcing is added to the system the fluctuation-dissipation theorem is recovered (Marini Bettolo Marconi et al. 2008). Moreover, some papers have shown that a direct application of the fluctuation dissipation theorem in a climatic context is reasonably successful (see, *e.g.*, Gritsun and Branstator 2007). It is definitely worth exploring whether this results exactly from the fact that numerical schemes introduce at all practical effects noise into the climate models, or from the fact that in the specific case of climate in present conditions the violation of the fluctuation-dissipation theorem is numerically small. A recent paper by Wouters and Lucarini (2013) suggests that, actually, the fluctuation-dissipation relation may be usable also in the case of non-equilibrium steady state systems if one focuses on coarse grained properties.

Non-equilibrium statistical mechanics also provides exciting tools for defining new strategies for the understanding of basic processes involved in large-scale climate dynamics, including also feedback mechanisms, and for treating rigorously ensembles of model simulations. The Ruelle response theory and its extension in the frequency domain allow the formulation of a new way of studying rigorously, the response of the climate system to perturbations, and to provide the foundation for defining what we may call the *spectroscopy of the climate system*, which provides the possibility of evaluating, using a perturbative approach, climate sensitivity and climate change from a radically new perspective. This paves the way for studying a potentially immense class of problems.

## 7 Acknowledgements

VL wishes to thank G<sup>3</sup> for providing food for thought.

## References

- Allen, M. (1999). Do-it-yourself climate prediction. *Nature* 401. 642.
- Boschi, R., Lucarini V., & Pascale, S. (2013). Bistability of the climate around the habitable zone: a thermodynamic investigation. *Icarus*. DOI: 10.1016/j.icarus.2013.03.017.
- Eckmann, J.-P. & Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics* 57. 617–656.
- Fraedrich, K. (1979). Catastrophes and Resilience of a zero-dimensional climate system with ice-albedo and greenhouse feedback. *Quarterly Journal of the Royal Meteorological Society* 105. 147–167.



- Fraedrich, K., Jansen, H., Kirk, E., Luksch, U., & Lunkeit, F. (2005). The Planet Simulator: Towards a user friendly model. *Meteorologische Zeitschrift* 14. 299–304.
- Gallavotti, G. (2006). Nonequilibrium statistical mechanics (stationary): overview. In: Françoise, J.-P., Naber, G. L., & Tsou Sheung Tsun (eds.). *Encyclopedia of Mathematical Physics*. Amsterdam: Elsevier. 530–539.
- Gritsun, A. & Branstator, G. (2007). Climate response using a three-dimensional operator based on the fluctuation-dissipation theorem. *Journal of the Atmospheric Sciences* 64. 2558–2575.
- Groot, S. R. de & Mazur, P. (1962). *Nonequilibrium Thermodynamics*. Amsterdam: North-Holland.
- Held, I. M. (2005). The Gap between Simulation and Understanding in Climate Modeling. *Bulletin of the American Meteorological Society* 86. 1609–1614.
- Hasselmann, K. (1976). Stochastic climate models. Part 1: Theory. *Tellus* 8. 392–400.
- Johnson, D. R. (2002). Entropy, the Lorenz Energy Cycle and Climate. In: Randall, D. A. (ed.). *General Circulation Model Development: Past, Present and Future*. New York: Academic Press. 659–720.
- Kleidon, A. & Lorenz, R. D. (eds.) (2005). *Non-equilibrium Thermodynamics and the Production of Entropy. Life, Earth, and Beyond*. Berlin: Springer.
- Kubo, R. (1966). The fluctuation-dissipation theorem. *Reports on Progress in Physics* 29. 255–284.
- Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., & Schellnhuber, H. J. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences* 105. 1786–1793.
- IPCC (2007). *The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
- Lorenz, E. N. (1967). *The Nature and Theory of the General Circulation of the Atmosphere*. Geneva: World Meteorological Organization.
- Lorenz, E. N. (1979). Forced and free variations of weather and climate. *Journal of the Atmospheric Sciences* 36. 1367–1376.
- Lucarini, V. (2008a). Validation of Climate Models. In: Philander, G. (ed.), *Encyclopedia of Global Warming and Climate Change*. Thousand Oaks: SAGE. 1053–1057.
- Lucarini, V. (2008b). Response Theory for Equilibrium and Non-Equilibrium Statistical Mechanics: Causality and Generalized Kramers-Kronig relations. *Journal of Statistical Physics* 131. 543–558.
- Lucarini, V. (2009a). Thermodynamic Efficiency and Entropy Production in the Climate System. *Physical Review E* 80. 021118.
- Lucarini, V. (2009b). Evidence of dispersion relations for the nonlinear response of the Lorenz 63 system. *Journal of Statistical Physics* 134. 381–400.
- Lucarini, V., Calmanti, S., Dell'Aquila, A., Ruti, P. M., & Speranza, A. (2007). Intercomparison of the northern hemisphere winter mid-latitude atmospheric variability of the IPCC models. *Climate Dynamics* 28. 829–849.
- Lucarini, V., Fraedrich, K., & Lunkeit, F. (2010a). Thermodynamic Analysis of Snowball Earth Hysteresis Experiment: Efficiency, Entropy Production, and Irreversibility. *Quarterly Journal of the Royal Meteorological Society* 136. 2–11.
- Lucarini, V., Fraedrich, K., & Lunkeit, F. (2010b). Thermodynamics of Climate Change: Generalized Sensitivities. *Atmospheric Chemistry and Physics* 10. 9729–9737.

- Lucarini, V., Fraedrich, K., & Ragone, F. (2011). New results on the thermodynamical properties of the climate system. *Journal of the Atmospheric Sciences* 68. 2438–2458.
- Lucarini, V. & Sarno, S. (2011). A Statistical Mechanical Approach for the Computation of the Climatic Response to General Forcings. *Nonlinear Processes in Geophysics* 18. 7–28.
- Marini Bettolo Marconi, U., Puglisi, A., Rondoni, L., & Vulpiani, A. (2008). Fluctuation-dissipation: response theory in statistical physics. *Physics Reports* 461. 111–195.
- Navarra, A., Kinter III, J. L., & Tribbia, J. (2010). Crucial Experiments in Climate Science. *Bulletin of the American Meteorological Society* 91. 343–352.
- Palmer, T. & Williams, P. (eds.) (2010). *Stochastic physics and climate modelling*. Cambridge: Cambridge University Press.
- Pascale, S., Gregory, J. M., Ambaum, M., & Tailleux, R. (2009). Climate entropy budget of the HadCM3 atmosphere ocean general circulation model and of FAMOUS, its low-resolution version. *Climate Dynamics* DOI:10.1007/s00382-009-0718-1.
- Peixoto, J. & Oort, A. (1992). *Physics of Climate*. New York: Springer.
- Ozawa, H., Ohmura, A., Lorenz, R., & Pujol, T. (2003). The second law of thermodynamics and the global climate system: a review of the maximum entropy production principle. *Reviews of Geophysics* 41. 1018.
- Ruelle, D. (1998). General linear response formula in statistical mechanics, and fluctuation-dissipation theorem far from equilibrium. *Physics Letters A* 245. 220–224.
- Ruelle, D. (2009). A review of linear response theory for general differentiable dynamical system. *Nonlinearity* 22. 855–870.
- Saltzman, B. (2002). *Dynamic Paleoclimatology*. New York: Academic Press.
- Schertzer, D. & Lovejoy, S. (2004). Uncertainty and Predictability in Geophysics: Chaos and Multifractal Insights. In: Sparks, R. S. J. & Hawkesworth, C. J. (eds.). *State of the Planet, Frontiers and Challenges in Geophysics*. Washington: AGU. 317–334.
- Shukla, J., Hagedorn, R., Kinter, J., Marotzke, J., Miller, M., Palmer, T., & Slingo, J. (2009). Revolution in climate prediction is both necessary and possible: A declaration at the World Modelling Summit for Climate Prediction. *Bulletin of the American Meteorological Society* 90. 175–178. DOI:10.1175/2008BAMS2759.1.
- Speranza, A. & Lucarini, V. (2005). Environmental Science: physical principles and applications. In: Bassani, F., Liedl, J., & Wyder, P. (eds.). *Encyclopedia of Condensed Matter Physics*. Amsterdam: Elsevier.
- Speranza, A., Buzzi, A., Trevisan, A., & Malguzzi, P. (1985). A theory of deep cyclogenesis in the lee of the Alps. Part I: Modifications of baroclinic instability by localised topography. *Journal of the Atmospheric Sciences* 42. 1521–1535.
- Tailleux, R. (2010). Entropy versus APE production: On the buoyancy power input in the oceans energy cycle. *Geophysical Research Letters* 37, L22603.
- Wouters, J. & Lucarini, V. (2012). Disentangling multi-level systems: averaging, correlations and memory. *Journal of Statistical Mechanics* P03003. DOI:10.1088/1742-5468/2012/03/P03003.
- Wouters, J. & Lucarini, V. (2013). Multi-level Dynamical Systems: Connecting the Ruelle Response Theory and the Mori-Zwanzig Approach. *Journal of Statistical Physics*. DOI: 10.1007/s10955-013-0726-8.
- Wunsch, C. & Ferrari, R. (2004). Vertical mixing, energy, and the general circulation of the oceans. *Annual Review of Fluid Mechanics* 36. 281–314.

**Prof. Valerio Lucarini**

University of Hamburg  
Meteorological Institute  
Grindelberg 5  
20144 Hamburg  
Germany  
valerio.lucarini@zmaw.de



Gregor Betz

# Chaos, Plurality, and Model Metrics in Climate Science

Commentary on Valerio Lucarini

## 1 Central findings of climate science are independent of model simulations

Since the reliability of climate models represents a politically highly sensitive issue, I would like to remind us upfront, before I comment on the interesting and illuminating paper by Valerio Lucarini, that many central findings of climate science are entirely independent of Global Climate Models (GCMs). These results include:

1. The atmospheric CO<sub>2</sub>-concentration has reached levels unprecedented in at least the past 650,000 years (IPCC 2007, p. 24). More specifically, the CO<sub>2</sub>-concentration varied, during the ice age cycles, between 180 and 300 ppm (IPCC 2007, p. 435).
2. The increase of atmospheric CO<sub>2</sub> from a pre-industrial concentration of 280 ppm to 380 ppm in 2005 is caused by human activities, notably by the consumption of fossil fuels (IPCC 2007, p. 25).
3. CO<sub>2</sub> is a greenhouse gas. Absorbing infrared light, it contributes to the natural greenhouse effect that heats the earth – as well as the planet Venus (Rahmstorf and Schellnhuber 2006, p. 32).
4. Increasing the CO<sub>2</sub>-concentration is a major intervention into the global climate system, offsetting the earth's radiative equilibrium and thus causing major readjustments of the climate system. These readjustments might consist in global warming or an increase of the earth's albedo.
5. Global average surface temperature has increased by roughly 0.6° during the last century. In the Polar Regions, where climate change, as a consequence of the ice albedo feedback, is expected to be more severe, surface temperature has been increasing at twice the rate of the rest of the world. (IPCC 2007, p. 37)
6. Except for CO<sub>2</sub>, known forcings of the climate system exhibit no trend during the last decades of the 20<sup>th</sup> century. Therefore, at least the latest phase of observed global warming can be attributed to anthropogenic activities. (Rahmstorf and Schellnhuber 2006, pp. 39–40)

These findings alone might constitute a sufficient reason for considering climate change a serious global problem, which has to be addressed by suitable policies. That is why even a radical criticism of GCMs does not automatically lend support to the position of so-called climate sceptics, i.e. the position that the theory of anthropogenic climate change is simply made-up and does not call for any policy measures whatsoever.

This said, reliable Global Climate Models would nevertheless be highly valuable for practical matters, as Lucarini has rightly stressed, because there are some things we apparently cannot estimate without GCMs.

## **2 Some relevant questions cannot be answered without GCMs**

GCMs are required to specify

1. the precise extent and timing of future global warming;
2. the regional patterns of future temperature and precipitation change;
3. the precise degree to which human activity is responsible for already observed climate change;
4. the detailed reconstruction of past climates from (sparse) proxy data.

With regard to the rational deliberation of alternative climate policy decisions, well-founded conditional predictions corresponding to the items 1 and 2 would obviously be extremely helpful. But can we reliably predict the climate? I understand that Lucarini cites three different uncertainties which prevent us from making accurate deterministic forecasts: ignorance about the precise initial conditions, ignorance about future boundary conditions, and ignorance about the causal structure of the climate system, which corresponds – in climate science jargon – to “structural uncertainty”. In his assessment of these uncertainties, Lucarini seems to presuppose that the climate system exhibits sensitive dependence on initial conditions. This prompts my first critical question.

## **3 Is the climate chaotic?**

Is the climate system chaotic; or, indeed, does an error in initial conditions grow exponentially when predicting the evolution of the climate system? This question, I suggest, deserves a careful and differentiated consideration. Granted: The

weather is chaotic. But this does not entail that the climate, which is described in terms of *average* weather, depends sensitively on initial conditions, too. It seems to me an obvious fact that some physical systems are chaotic with regard to the microstates they realize, but behave non-chaotically regarding their macrostates (Think, e.g., of boiling water, which is, regarding the location of the first vapour bubble, chaotic, yet is not in terms of the mean temperature when bubbles start to form.). By the way, this is maybe the very reason why reduction of complexity (through devising highly aggregated models) can represent a successful research strategy. So, even if the weather is chaotic, the climate is not necessarily so, and in particular not necessarily with regard to all its state variables. As a philosopher, I am, of course, not in a position to answer the empirical question which climate variables depend sensitively on initial conditions. But I would like empirical scientists to be more specific regarding the chaos hypothesis. Here is a suggestion for how the chaotic character of the climate system might be described in a more nuanced way:

- Some large-scale climate processes such as the thermohaline circulation or the indian monsoon possess, under specific boundary conditions, several equilibria. In such situations, small perturbations might determine whether the respective subsystem ends up in one or the other stable state. These climatic changes might be “abrupt” and trigger global effects (affecting, e.g. global precipitation or temperature patterns).
- Whether, say, average precipitation in northern Germany in the decade 2100–2110 is going to be higher or lower than in the current decade (2000–2010) possibly also depends on the precise climatic initial and boundary conditions such as today’s radiative forcing, heat uptake of the ocean, state and interplay of atmospheric oscillations, etc.
- But whether the emission of another 1000 GtC in the first half of this century causes the earth to warm, in 2100, by 10 or by 2 degree Celsius does not depend sensitively on today’s initial conditions.

## 4 The plurality of GCMs

Of the three key uncertainties Lucarini enumerated, namely (1) ignorance of initial conditions, (2) ignorance of boundary conditions, and (3) structural uncertainty, it is the last one which is responsible for the plurality of climate models employed in climate science. Unlike in economics, however, climate scientists do understand the basic, small-scale processes in the climate system. The fundamental laws describing these processes, such as the Navier-Stokes equation, are well estab-

lished. It is only because of limited computational resources that these equations cannot be solved for a system as huge as earth's climate. The computational limitations call for a description of climate processes on a more aggregate scale – and it is precisely on this meso-scale where the causal picture of the climate is still inadequate. When devising a GCM, climate scientists face, as a consequence, a couple of underdetermined choices, and different groups of modellers end up with different climate models (cf. Parker 2006; Betz 2009; Lenhard and Winsberg 2010).

Thus, the 4AR relies for its major predictions as well as for the analysis of past climates on 23 different AOGCMs which are built and run by 17 institutions (IPCC 2007, p. 597). These GCMs comprise sub models of the atmosphere, the ocean, sea ice and land. Their resolutions range from  $1,1^{\circ} \times 1,1^{\circ} - 4^{\circ} \times 5^{\circ}$  with 56–12 vertical layers (where, at the equator, one degree of latitude equals a degree of longitude and amounts, roughly, to 111 km).

Given this plurality of models, the question whether one can empirically test, compare and rank these rival models arises quite naturally. This question will eventually lead us to one of Lucarini's main points, namely the proposal of a new metric for climate model evaluation.

## 5 Epistemic evaluation of GCMs: the role of metrics

Regarding the epistemic assessment of GCMs, it is important to separate the following two questions:

- (1) What are the empirical implications of a climate model that ought to be considered during its epistemic assessment at all?
- (2) What exactly can one infer from the predictive and explanatory performance of a GCM regarding the relevant empirical indicators?

The second question pertains to the general methodology of the model assessment: Should we try to falsify GCMs and refute those that, for example, give rise to false empirical retrodictions? Or do we have to construe the model evaluation along the lines of inductive modes of reasoning? Or should one assess the models in agreement with a hypothetico-deductive account of confirmation? – We will return to this second question below.

The first question concerns a more rudimentary issue, which has to be addressed before any kind of empirical assessment can be carried out. An answer to the first question, which somehow suggests itself, is to say: The empirical implication *E* is relevant if and only if *E* concerns climate variables the model is supposed to predict or to explain. And that is roughly what the IPCC assumes



(IPCC 2007, chapter 8). Accordingly, the aspects of the climate system considered in course of the model evaluation include regional mean surface temperature, the annual variability of surface temperatures, mean and annual variability of precipitation, patterns of cyclone activities, mean temperature and salinity structure of the ocean, strength and geometry of ocean circulations, the extent of sea ice, the severity and frequency of extreme weather events, large-scale processes such as the monsoon or El Niño, etc.

This plurality of relevant climate variables poses a potential problem for the assessment of GCMs since there is no climate model which outperforms its rivals in terms of empirical adequacy and with regard to all the different relevant aspects of the climate system. Every model has some strengths and some deficiencies, and they typically differ from the strengths and deficiencies of its rivals (Heffernan 2010).

It is in this situation that climate scientists would like to devise a general quantity, which aggregates all the relevant aspects, and which allows one to express the overall empirical adequacy of a GCM in one single figure. Such an aggregated variable is also referred to as a metric. The IPCC defines a metric as a consistent measurement of an object's or activity's characteristic that is otherwise difficult to quantify. (IPCC 2007, p. 949)

In its Third Assessment Report (TAR), published in 2001, the IPCC was unambiguous about any attempts to design a metric that combines all relevant empirical implications of a GCM:

It has proved elusive to derive a fully comprehensive multi-dimensional “figure of merit” for climate models. (IPCC 2001, p. 475)

In the 4AR, however, the IPCC has become a bit more optimistic, again:

The possibility of developing model capability measures (‘metrics’), based on the above evaluation methods, that can be used to narrow uncertainty by providing quantitative constraints on model climate projections, has been explored for the first time using model ensembles. While these methods show promise, a proven set of measures has yet to be established. (IPCC 2007, p. 60)

On this background, Lucarini proposes his own, *process-oriented* metrics for model evaluation. As far as I understand, these metrics do not rely on policy-relevant observational variables but try, rather, to capture the key processes of the climate system. They are supposed to describe, based on simulation results or on observational data, the central causal mechanisms that drive the (simulated or real) climate system – what Lucarini also calls the “climatic machine”. This would enable us to differentiate, for example, between (a) GCMs that perform

well in regard of the reproduction of policy-relevant observational trends but do not get the underlying causal mechanisms right and (b) GCMs with a fairly good representation of key climate processes but a poor performance in terms of policy-relevant variables.

As a philosopher, I cannot judge whether Lucarini's proposal is suited to capture some key climate processes. So that is something I take for granted in the following discussion.

Lucarini's proposal raises the interesting question how the process-oriented metrics relate to the traditional ones based on fields of practical interest. I take it that Lucarini does not mean to replace traditional metrics by one or several process-oriented ones. Still, I see a bunch of questions that deserve further discussion:

- Should the process-oriented metrics be the primary indicator for the reliability of the predictions of a GCM? So, e.g., does the empirical inadequacy of a GCM in terms of a process-oriented metric undermine the credibility of its policy-relevant predictions, even if the model performs well in terms of those policy-relevant variables?
- What are the underlying assumptions that justify the expectation that the improvement of models in terms of process-oriented metrics leads, in the long run, to more accurate predictions in terms of policy-relevant variables?

These points inevitably lead us back to the question what at all one may infer from a good or bad performance of a GCM in terms of the relevant variables; that is back to the second question stated above.

## 6 Interpreting multi-model ensembles

Every GCM has false empirical implications. This holds for policy-relevant implications as well as, I assume, for process oriented metrics. According to a falsificationist methodology, all GCMs would have to be rejected. The ensemble of GCMs the IPCC relies on can thus not be understood as comprising all models not yet falsified. Falsificationism is of no help to understand the status of GCMs and the way they are assessed given their empirical performance.

Lucarini asserts that the model ensembles must not be interpreted probabilistically, either. (Yet, he seems to be ambiguous on this point, claiming in the concluding section that climate results “must be plural and stated in probabilistic terms”.) I wholeheartedly agree that the empirical implications of GCMs, as of today, cannot be used to derive a probabilistic interpretation of the model ensemble.

ble (see also Betz 2007; Parker 2010). Let me briefly sketch why: (1.) Assigning probabilities to different GCMs only makes sense in a subjectivist interpretation of probabilities. (2.) Climate data does not significantly constrain the posteriors, which still depend crucially on the prior probabilities. (3.) These priors are really arbitrary, because climate scientists do not possess sufficient tacit knowledge (of 21<sup>st</sup> century climate change) to justifiably constrain the priors. Probabilistic studies in climatology rely on arbitrary, typically uniform priors. In particular, the catch-all hypothesis that a model not yet devised provides a correct analysis, is typically assigned the value zero.

This said, what does an ensemble of GCMs tell us? A group of climatologists from the Hadley Centre, who basically share the above diagnosis, have put forward an interesting proposal. In an article published in 2007, Stainforth et al. suggest:

Today's ensembles give us a lower bound on the maximum range of uncertainty. (Stainforth, Allen et al. 2007, p. 2156)

So, in other words, whatever happens in a model simulation might actually happen in the future. Still, the future evolution of the climate system might also follow a dynamic that is not predicted by any GCM yet. That is, the range of possible evolutions of the climate system, given our current understanding, comprises *at least* the predictions of the model ensemble.

This is arguably a very modest interpretation of the epistemic status of GCMs. I would say, however, that this is the correct interpretation.

One might wonder whether, according to this interpretation of model ensembles, climate models are assessed in terms of their empirical implications at all. Is not the methodological outlook of Stainforth et al. flawed, or at least incomplete, as long as it does not explain how GCMs are evaluated on the basis of relevant climate data? We can address this challenge as follows: The empirical data regarding relevant climate variables already enters the process of constructing climate models. Model versions that perform definitely very poorly are excluded by so-called tuning. It is the calibration of model parameters that makes sure that only GCMs with a comparatively high empirical adequacy enter the model ensemble. And maybe it is exactly here, in the calibration process, where Luca-rini's process-oriented metrics might have a major role to play. But that remains a further open question.

## 7 Improving our epistemic situation

As a final remark, I would like to draw our attention to the question how to construe scientific advancement given the specific interpretation of GCMs. Obviously, climate scientists should keep on trying to establish reliable and justified probability forecast, and succeeding in doing so would count as a major scientific breakthrough. But provided these attempts fail and the model ensemble remains merely a lower bound on the range of uncertainty, what sorts of changes count as improvement of our epistemic situation? What does scientific progress, in such a situation, mean at all?

Counter-intuitively, progress might consist in widening the range of models and their predictions. This could be achieved through devising ever new models, for instance by systematically varying all uncertain model assumptions and including ever new, relevant processes into the models. And that is, partly, what happens in the climate community. According to a recent News Feature in *Nature*, some climate scientists expect that this process will lead to a significant extension of the span of climate predictions, proving the ranges of previous IPCC reports too narrow.

It's very likely that the generation of models that will be assessed for the next IPCC report will have a wider spread of possible climate outcomes as we move into the future. (Jim Hurrell, National Center Atmospheric Research in Boulder, Colorado, quoted in (Heffernan 2010, p. 1014))

Given the interpretation of model ensembles by Stainforth et al., such an extension of the scenario range is nothing climate scientists should fear – but rather strive for.

## References

- Betz, G. (2007). Probabilities in Climate Policy Advice: A Critical Comment. *Climatic Change* 85(1–2). 1–9.
- Betz, G. (2009). Underdetermination, Model-ensemble, and Surprises – On the Epistemology of Scenario-analysis in Climatology. *Journal for General Philosophy of Science* 40(1). 3–21.
- Heffernan, O. (2010). The Climate Machine. *Nature* 463(7284). 1014–1016.
- IPCC (2001). *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

- IPCC (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge; New York: Cambridge University Press.
- Lenhard, J. & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics* 41(3). 253–262.
- Parker, W. S. (2006). Understanding Pluralism in Climate Modeling. *Foundations of Science* 11. 349–368.
- Parker, W. S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Modern Physics* 41(3). 263–272.
- Rahmstorf, S. & Schellnhuber, H. J. (2006). *Der Klimawandel*. München: C. H. Beck.
- Stainforth, D. A., Allen, M. R., et al. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society A – Mathematical Physical and Engineering Sciences* 365(1857). 2145–2161.

**Jun.-Prof. Dr. Gregor Betz**

Karlsruhe Institute of Technology (KIT)

Institute of Philosophy

Kaiserstraße 12

Geb. 20.12

76131 Karlsruhe

Germany

gregor.betz@kit.edu



# Subject Index

- action 9, 10, 19, 36, 37, 119, 124, 127, 128, 131, 132, 133, 134, 135, 141, 142, 145, 146, 147, 148, 195, 229
- ad hoc modifications 185
- analogy 48, 72, 93, 128, 129, 130, 131, 132, 133, 134, 142
- anomaly 24
- a priori knowledge 190, 191, 197, 222, 223, 225
- audience 89, 91, 92, 93, 94, 107, 108
  
- balances 188, 193, 199, 202, 205, 207, 230, 237, 243
- bootstrap model of confirmation 25
  
- causality problem 22, 246
- chemical engineering 7, 187, 201, 208, 212, 221, 222
- climate change 7, 232, 236, 241, 242, 243, 246, 249, 250, 255, 256, 261
- climate models 7, 85, 232, 233, 234, 235, 236, 237, 238, 239, 240, 246, 248, 250, 255, 256, 257, 258, 259, 261
- climate projections 233, 247, 259
- climate sceptics 256
- climate variability 232, 236, 249
- complex 1, 2, 4, 5, 6, 44, 48, 53, 55, 56, 62, 63, 64, 76, 81, 87, 88, 94, 95, 103, 109, 123, 125, 131, 134, 143, 153, 155, 173, 175, 179, 184, 185, 187, 188, 197, 206, 208, 209, 211, 212, 222, 229, 230, 231, 232, 241, 243
- computer simulation 1, 71, 87, 88
- computing 196, 233, 236, 246, 249
- consistency 21, 223, 236, 246
- cortical dissociation 120
- cosmic microwave background (CMB) 3, 11, 14, 24
- cosmological constant 18, 19, 20
- cosmological term 18
- coupling 33, 34, 37, 38, 65, 124, 135, 232, 233
  
- dark energy 3, 18, 19, 20, 21, 23, 26, 71
- data-driven 89
  
- diachronic coherence 238
- diffusion in liquids 7, 189, 200, 206
- direction of fit 221
- Duhem-Quine problem 89, 107
- dynamical systems 160, 230, 243
  
- efficiency 17, 241, 244, 245, 249
- energy balance 244
- ensemble 55, 236, 238, 239, 240, 260, 261, 262
- entropy 232, 244, 245, 249
- error-in-variables estimation 199, 208
- error-in-variables method 199
- evaluative function 4, 23
- evolutionary novelty 47
- exoplanets 249
- experimental validation 204, 206, 208, 210
- explorative function 4, 23
- extremal principle 9, 10, 11, 19
  
- falsification 26, 169, 170, 180, 181, 238
- falsificationism 160, 180, 181, 223
- fluctuation-dissipation theorem 246, 249, 250
- functional definition 121, 130, 131
- functional restructuring 121
  
- gaits 32, 33, 34, 35, 39
- geo-engineering 241, 242, 243
- global warming 241, 255, 256
- goodness of fit 6, 75, 153, 162, 164, 166, 167, 168, 171, 174, 179, 181, 182
  
- hallucinations 4, 36, 38, 39
- horizontal comparison 239
- Hubble constant 13
  
- identifiability analysis 190, 199, 212
- inflationary expansion 22, 27
- interplay of experimental results and theoretical predictions 238
- inverse problem 188, 190, 191, 198, 200, 202, 209, 210, 211
- irreversible processes 232

- isolation 93, 100, 107
- isotropy 11, 15, 16, 19, 20, 21, 26
- iterative model identification strategy 192
- iterative refinement strategy 189
- kinetics 7, 189, 195, 197, 200, 201, 205, 206, 212, 213, 221, 224, 225, 226
- lesion 120
- low-frequency variability 243
- mathematical model 4, 29, 39, 43, 44, 45, 47, 48, 49, 53, 55, 56, 74, 87, 163, 187, 189, 190, 191, 213, 221, 222
- mathematical models 44
- maximum likelihood approach 196, 197
- metric 10, 240, 258, 259, 260
- model commentary 93, 94, 97, 101
- model description 91, 92, 107
- model discrimination 195, 198, 199
- model-independent 25
- modeling risks 98
- model resolution 189
- molecular structure 7, 223, 224, 225
- natural selection 46, 47
- negative evidence 24
- neuronal selectivity 123, 124
- non-equilibrium statistical mechanics 232, 243, 246, 249, 250
- non-equilibrium system 232, 246
- normative approach 123, 124, 125, 126, 127
- numerical methods 191, 200, 230
- object identification 122, 126
- object recognition 122, 123, 124, 125, 126, 127, 128, 130, 131, 132, 134, 135, 141, 142, 143
- observables 236, 237
- optimal design of experiments 190, 191, 198, 200, 210, 212
- organismal form 43, 45, 47, 48, 49
- parameterization 199, 207, 233, 237, 244, 247
- parameter precision 197, 198, 199, 204
- phase transitions 59, 63, 64, 72, 232
- plurality 4, 7, 46, 255, 257, 258, 259
- portable model structures 88
- predictive analogies 128, 129, 130, 131, 135
- principled division 117, 118, 120, 129
- probability 61, 84, 154, 157, 160, 167, 168, 171, 172, 174, 175, 240, 246, 262
- process-oriented metrics 7, 237, 238, 259, 260, 261
- projections 238
- qualitative methods 166
- quality of fit 199, 203
- range of possible evolutions of the climate system 261
- reaction mechanism 7, 223, 224, 225, 226
- reciprocal synchronic coherence 238
- reductionist scientific method 230
- redundancy reduction 123
- reliability of data 6, 153, 166, 167, 174, 179, 180
- resemblance 12, 91, 92, 93, 94, 98, 100, 101, 103, 107, 109, 181, 248
- resolution 188, 191, 192, 193, 196, 198, 200, 201, 204, 206, 209, 233, 234, 235, 237, 258
- response properties 123, 124, 125
- reverse modeling 173
- Ruelle response theory 246, 250
- scientific advancement 262
- sensory processing 117, 118, 119, 121, 123, 124, 126
- simple models 1, 56, 59, 60, 61, 62, 64, 82, 95
- simultaneous model identification (SMI) strategy 192, 196, 223
- slowness 124, 126, 127
- spatio-temporal symmetries 33, 34
- stability 29, 30, 31, 123, 124, 125, 127, 141, 174, 239, 243, 245
- stochastic forcing 243, 250
- strict laws 44
- structural properties 121, 231
- structural uncertainties 239, 248
- structure-mapping theory 128
- sub-models 195, 196, 198, 211



- substance concept 143, 144, 148, 149
- surrogate object 87, 90, 91, 94, 97, 98, 100, 104
- surrogate reasoning 87, 88, 94
- symmetry 3, 4, 10, 11, 12, 15, 19, 20, 21, 30, 31, 33, 34, 37, 38, 39, 43, 45, 48, 49
- symmetry-breaking 4, 29, 30, 36, 37, 39, 43, 45, 48, 49
- target object 87, 91, 92, 93
- temporal coherence 124, 127
- temporal contiguity 125
- theory-driven 88
- thermodynamics 7, 230, 232, 243, 245, 249
- thin conception 225
- time scale(s) 13, 17, 55, 124, 188, 230, 232, 243, 246, 248
- tipping point(s) 64, 72, 232, 245, 248, 249
- tractability 95, 103
- transport in wavy falling film flows 189
- unexpected predictions 153, 166, 168, 174, 179, 181
- unrealisticness 89, 96
- validation 67, 236, 237, 248, 249
- verification 168, 236
- vertical comparison 239
- visual definition 130, 131
- world-experiments 248



# Author Index

- Acemoglu, D. 112  
Adomeit, P. 209  
Agarwal, M. 195  
Akaike, H. 199, 204  
Akerlof, G. 89, 111, 112  
Albeverio, S. 63  
Allen, M. 249, 261  
Alsmeyer, F. 206  
Amrhein, M. 201  
Anderson, D. R. 189, 199, 208  
Arsenin, V. Y. 202  
Asprey, S. P. 191  
Asselt, M. B. A. van 75
- Bach, J. 119  
Balakotaiah, V. 208  
Barbas, H. 121  
Bard, Y. 189, 196, 197, 203, 213  
Bardow, A. 187, 188, 189, 199, 201, 202, 206, 207, 208, 211, 214  
Barker, F. G. 120  
Barlow, H. B. 123  
Barsalou, L. W. 142  
Bartelmann, M. VI, 3, 9, 22, 23, 24, 25, 26  
Bastin, G. 212  
Beck, J. V. 213  
Beer, R. D. 160  
Bell, A. J. 124  
Bentley, R. A. 73  
Berkes, P. 125  
Bertalanffy, L. von 55  
Betz, G. VI, 7, 255, 258, 261, 263  
Bhatt, N. 201  
Biddle, J. 114  
Biegler, L. T. 189, 197  
Bird, R. B. 188, 206  
Bischof, C. 212  
Bishop, C. M. 126  
Blaisdell, A. P. 118  
Bonner, J. T. 47, 48  
Bonvin, D. 203  
Bothe, D. 206  
Box, G. E. P. 60, 170, 175  
Braitenberg, V. 119
- Brakel, J. van 221, 222  
Brandon, R. N. 48  
Branstator, G. 250  
Bremmer, F. 146  
Brendel, M. 198, 200, 201, 204, 214  
Bressloff, P. C. 36, 37, 38, 45  
Britt, H. I. 199, 208  
Brockfeld, E. 67  
Buchwald, J. 222  
Buono, P.-L. 34  
Burnham, K. P. 189, 199, 208  
Buss, L. W. 46
- Calenbuhr, V. 72  
Carpenter, B. 156, 223  
Cartwright, N. 1, 3, 82, 114  
Cassidy, J. 112  
Cheng, Z. M. 196  
Chumbley, J. I. 163  
Colander, D. 101, 108, 112  
Collins, J. J. 33  
Comte, A. 53  
Cowan, J. D. 36, 37, 38  
Cowey, A. 120  
Craighero, L. 145
- Darwin, C. R. 46  
Davidson, P. A. 62  
DiCarlo, J. J. 125  
Dobzhansky, T. 44  
Dochain, D. 212  
Dostrovsky, J. 125  
Douglas, R. J. 121  
Draper, N. R. 60  
Drieghe, D. 157  
Dubois, D. 75
- Eckmann, J.-P. 243  
Einhäuser, W. 125, 126  
Einstein, A. 3, 10, 12, 17, 19, 20, 21, 60  
Elster, J. 113  
Engbert, R. VI, 6, 153, 154, 155, 158, 159, 161, 162, 163, 164, 169, 170, 171, 173, 175, 178, 179, 180, 181, 182, 183, 184, 185

- Engl, H. W. 198, 199, 200, 201, 207  
 Epstein, J. M. 57, 61, 63  
 Erdogan, S. T. 146  
 Essen, D. C. van 121, 123, 144, 146  
 Evans, T. 128  
 Everling, S. 173
- Fadiga, L. 146, 148  
 Falkenhainer, B. 128  
 Farah, M. J. 120  
 Farmer, J. 89  
 Fehr, E. 74  
 Felleman, D. J. 121, 123, 144  
 Fernald, R. D. 117  
 Ferrari, R. 245  
 Ferreira, F. 157, 161  
 Festa, R. 84  
 Field, M. 44, 49, 124, 125  
 Findlay, J. M. 156  
 Fisher, D. L. 156  
 Floreano, D. 63  
 Földiák, P. 124  
 Fontana, W. 46  
 Fraassen, B. C. van 1, 2, 179  
 Fraedrich, K. 232, 249  
 Franceschini, G. 197  
 Franzius, M. 125, 126  
 Frege, G. 130  
 Frey, B. 74  
 Frigg, R. 1, 82  
 Froment, G. F. 212
- Gähde, U. V, VI, 1  
 Gallavotti, G. 246  
 Gallese, V. 146  
 Galton, F. 61, 75  
 Garrigues, P. 124  
 Gentner, D. 128  
 Gibson, J. J. 135, 142  
 Giere, R. 1, 84, 91  
 Gigerenzer, G. 74  
 Gilbert, N. 63  
 Gilbert, S. F. 46  
 Gintis, H. 73  
 Glymour, C. 25  
 Gödel, K. 68  
 Golub, G. H. 202
- Golubitsky, M. VI, 4, 29, 30, 31, 34, 37, 39, 42, 43, 44, 45, 48, 49  
 Goodwin, W. 225  
 Gould, S. J. 46, 48  
 Grabner, E. 153  
 Grinvald, A. 145  
 Gritsun, A. 250  
 Groot, S. R. de 232, 243  
 Grüne-Yanoff, T. 111  
 Guala, F. 63  
 Gust, H. 129
- Hacking, I. 1, 222  
 Hamilton, R. H. 9, 121  
 Hanke, M. 200  
 Hansen, P. C. 202  
 Hartmann, S. V, VI, 1, 2, 3, 5, 81, 82, 84, 86  
 Hasselmann, K. 243  
 Hastie, T. 195  
 Haxby, J. V. 123  
 Heffernan, O. 259, 262  
 Helbing, D. VI, 4, 5, 53, 54, 60, 61, 63, 64, 65, 67, 69, 71, 76, 79, 81, 82, 83, 84, 85  
 Held, I. M. 232, 248  
 Henderson, J. M. 157, 161  
 Hendry, R. F. VI, 7, 221, 223, 227  
 Hesse, M. 1  
 Higham, D. J. 195  
 Hilgetag, C. C. 121  
 Hindriks, F. 93, 103  
 Hipp, J. 125  
 Hodgson, G. 102, 112  
 Hoffmann, M. VI, 6, 179, 186  
 Hoffmann, R. 226  
 Hofstadter, D. 128  
 Holyoak, K. 129  
 Horsthemke, W. 59  
 Hosten, L. H. 212  
 Huang, C. 198, 202  
 Huber, L. 118  
 Huey, E. B. 174  
 Hummel, J. 129  
 Hung, C. P. 123  
 Hyvärinen, A. 124
- Indurkha, B. 129  
 Inhoff, A. W. 157

- Ishida, H. 146  
Iyengar, S. S. 213
- Jackson, M. 63  
Jacobs, A. M. 157  
Johnson, D. 232, 244  
Juselius, K. 113  
Just, W. 59
- Kaas, J. H. 117, 146  
Kagel, J. H. 63  
Kahneman, D. 74  
Karatashvili, M. 196, 199, 208, 209, 210, 214  
Kennedy, A. 163, 164  
Kerr, P. W. 156, 157  
Kesting, A. 67  
Keysers, C. 146  
Kietzmann, T. C. VI, 5, 117, 118, 139, 141  
Kim, W. 166  
Kirman, A. 112  
Kirsch, A. 188, 199  
Kitcher, P. 112, 114  
Kittrell, J. R. 191, 212, 213  
Kleidon, A. 232, 244, 245  
Klein, D. J. 20, 125  
Kliegl, R. VI, 6, 153, 155, 156, 157, 161, 162, 169, 171, 174, 178, 179, 180, 181, 182, 183, 184, 185  
Klipp, E. 200  
Knuuttila, T. 87  
Kokinov, B. 129  
König, P. VI, 5, 117, 124, 127, 139, 141, 142, 149  
Körding, K. P. 124, 125  
Körkel, S. 198  
Kriesten, E. 206, 208  
Krüger, N. 127, 141  
Krugman, P. 102, 103, 113  
Kubo, R. 246  
Kühnberger, K.-U. VI, 5, 117, 129, 139, 141  
Kuhn, T. S. 26, 68  
Kuorikoski, J. 88, 94
- Lakatos, I. 6, 26, 168, 171, 180, 181, 183, 184  
Lakoff, G. 146  
Land, M. F. 117
- Lawson, T. 110, 112, 113  
LeCun, Y. 126  
Lefever, R. 59  
Lehtinen, A. 88, 94  
Lenhard, J. 258  
Lenton, T. M. 232, 245  
Lerner, Y. 123  
Leuridan, B. 113  
Lewis, G. N. 226  
Lewis, J. W. 146  
Lewis-Williams, D. 36  
Lewontin, R. C. 46, 48  
Li, N. 125  
Lohmann, T. 203  
Longtin, A. 161  
Lorenz, J. 68, 85  
Lorenz, R. D. 232, 243, 244, 245, 246  
Lovejoy, S. 236  
Lovett, A. 130  
Lucarini, V. VI, 7, 75, 230, 232, 239, 244, 245, 246, 247, 248, 250, 253, 255, 256, 257, 258, 259, 260, 261
- Macchietto, S. 191, 197  
Mahoney, A. W. 212  
Mäki, U. VI, 5, 87, 88, 91, 93, 94, 96, 101, 103, 106, 107, 108, 109, 110, 111, 112, 113, 114  
Marchionni, C. 94  
Marini Bettolo Marconi, U. 250  
Markus, M. 212, 213  
Marquardt, W. VI, 6, 7, 187, 188, 189, 191, 192, 195, 198, 199, 200, 201, 202, 210, 211, 219, 221, 222, 223  
Martin, A. 148  
Martin, K. A. C. 121  
Mason, R. L. 190, 197  
Maye, A. 143, 146, 147  
Mayergoyz, I. D. 59  
Mayr, E. 44  
Mazur, P. 232, 243  
McCloskey, D. 113  
McConkie, G. W. 156, 157  
McDonald, S. A. 156, 162, 163, 164, 173  
McLure, M. 130  
Merabet, L. 121  
Meza, C. E. 208

- Mhamdi, A. 202, 214  
 Mihm, S. 112  
 Mikhailov, A. S. 72  
 Miller, J. H. 63  
 Millikan, R. G. 143  
 Mönnigmann, M. 213  
 Morgan, M. S. 3, 87, 88, 179, 185  
 Morrison, M. 3, 84, 179, 183, 185  
 Moussaid, M. 64, 73  
 Mroczko-Wąsowicz, A. VI, 148, 152  
 Munoz, D. P. 173  
 Musgrave, A. 93  
 Myrvold, W. 82  
 Myung, J. I. 166
- Navarra, A. 236, 248, 249  
 Navarro, D. J. 166  
 Neal, R. M. 175  
 Needham, P. 223  
 Neumann, J. von 103, 233  
 Niiniluoto, I. 84  
 Nilsson, D.-E. 117  
 Noe, A. 119  
 Nolfi, S. 63  
 Nuthmann, A. 153, 156
- O'Hara, R. J. 48  
 Oja, E. 124  
 O'Keefe, J. 125  
 O'Leary, D. P. 202  
 Olshausen, B. A. 124, 125  
 Oort, A. 229, 232  
 O'Regan, J. K. 119, 156  
 Ormerod, P. 73, 112  
 Ozawa, H. 244
- Page, S. E. 63  
 Parker, W. S. 258, 261  
 Pascale, S. 244  
 Pascual-Leone, A. 121  
 Pashler, H. 153, 162, 166, 167, 174, 179, 180, 184, 185  
 Peixoto, J. 229, 232  
 Petrov, A. 129  
 Pietronero, L. 71  
 Pinto, N. 44, 45, 118  
 Pitt, M. A. 165, 166
- Plessis, S. du 112  
 Pollatsek, A. 156, 157, 161  
 Pons, T. P. 146  
 Pope, S. B. 188  
 Popper, K. 65, 168, 180, 181, 191, 223  
 Prade, H. 75  
 Prausnitz, J. M. 195  
 Pronzato, L. 188, 190, 196, 197, 198, 199, 200, 203  
 Pukelsheim, F. 188, 190, 197, 198, 200  
 Pulvermüller, F. 146, 147, 148
- Quaiser, T. 190, 213  
 Quiggin, J. 98, 101  
 Quiroga, R. Q. 123
- Radach, R. 156, 161, 173  
 Rahmstorf, S. 255  
 Ramirez, W. F. 212  
 Ramsay, J. B. 213  
 Ramsay, J. O. 212, 213  
 Rao, M. S. 213  
 Rappelsberger, P. 148  
 Rayner, K. 155, 156, 157, 161, 163, 169  
 Reddix, M. D. 157  
 Reichle, E. D. 156, 157, 161, 162, 163, 164, 169, 173  
 Reilly, R. 156, 161, 173  
 Reiner, R. 48  
 Reinsch, C. H. 198, 207  
 Reiss, J. VI, 5, 111, 113, 114, 116  
 Renz, U. 209  
 Resnik, D. B. 48  
 Reydon, T. A. C. VI, 4, 43, 46, 51  
 Reymond, L. 117  
 Rice, S. H. 44  
 Richter, E. 156  
 Risse, S. 171, 175  
 Rizzolatti, G. 145, 146  
 Roberts, S. 153, 162, 166, 167, 174, 179, 180, 184, 185  
 Locke, A. 226  
 Rodríguez-Ferreiro, J. 148  
 Rolfs, M. 153  
 Romijn, R. 195  
 Roth, A. E. 63  
 Rotmans, J. 75

- Roubini, N. 112  
Rozzi, S. 146  
Ruelle, D. 232, 239, 243, 246, 249, 250  
Rumelhart, D. E. 126  
  
Sadato, N. 121  
Sahin, E. 146  
Sarno, S. 232, 246, 247  
Schad, D. J. 157, 161, 175  
Schagen, A. 210  
Schellnhuber, H. J. 255  
Schertzer, D. 236  
Schilling, H. E. H. 162, 163, 164, 167  
Schittkowski, K. 189, 197  
Schmidt, K. M. 74  
Schmidt, T. 204  
Schnitzler, A. 146  
Schuster, H. G. 59  
Schwering, A. 128  
Sejnowski, T. J. 124, 125  
Sen, A. 113  
Shillcock, R. C. 156  
Shiller, R. 89, 112  
Shmuel, A. 145  
Shukla, J. 248  
Singer, W. 146  
Sinigaglia, C. 146  
Slattery, J. C. 157, 188  
Smolin, L. 75  
Sneed, J. D. 179  
Sommerfeld, R. D. 21  
Speranza, A. 230, 247  
Sprekeler, H. 125  
Stainforth, D. A. 261, 262  
Stanley, H. E. 59, 63  
Steeves, J. 120  
Stegmüller, W. 179  
Stewart, I. 29, 30, 33, 37, 39, 43, 44, 45, 48, 49  
Stewart, W. E. 199, 208  
Stiglitz, J. 99, 100, 101, 112  
Stoerig, P. 120  
Strevens, M. 113  
Suárez, M. 179  
Sugden, R. 103, 114  
Suppes, P. 1, 179  
Sur, M. 121  
Swinburne, R. 82  
Sykes, P. 224, 226  
  
Tacca, M. C. VI, 5, 148, 152  
Tailleux, R. 245  
Tanaka, K. 123  
Taylor, R. 195, 213  
Teller, P. 109  
Thagard, P. 118  
Tholudur, A. 212  
Thompson, E. 47, 48, 49  
Thünen, J. H. von 96  
Tikhonov, A. N. 202  
Tilman, D. 66  
Timmer, J. 212  
Traulsen, A. 67  
Treiber, M. 67  
Trevelyan, P. M. J. 208  
Tversky, A. 74  
  
Vajda, S. 197  
Varela, F. 119  
Verheijen, P. J. T. 189, 190, 195  
Vespignani, A. 76  
Vicsek, T. 71  
Vitu, F. 156, 157  
Voss, H. U. 212  
  
Wahl, S. A. 196  
Walker, R. 156  
Walter, E. 188, 190, 196, 197, 198, 199, 200, 203  
Watanabe, S. 118  
Weidlich, W. 63  
Weisberg, M. 223  
Weiskrantz, L. 120  
Werbos, P. J. 126  
Werning, M. VI, 5, 143, 144, 146, 147, 148, 152  
Wilke, W. 209  
Williams, P. 233  
Wimsatt, W. 110  
Winsberg, E. 2, 114, 258  
Wiskott, L. 124, 125  
Woit, P. 75  
Wolf, J. H. V, VI  
Woodbury, K. A. 213

Wunsch, C. 245

Wyss, R. 125

Xiao, Y. 145

Ylikoski, P. 113

Yuan, W. K. 196

Zeeman, E. C. 29, 59, 63

Zeki, S. 120

Zihl, J. 120

Zola, D. 157

Zwaan, R. 134





