Towards a Theory of Thinking

Building Blocks for a Conceptual Framework

Britt M. Glatzeder
Vinod Goel
Albrecht von Müller

Editors

PARMENIDES BOOK SERIES
ON THINKING

PARMENIDES FOUNDATION

Springer

On Thinking

Series Editors

Ernst Pöppel
Parmenides Foundation, Kirchplatz 1, D-82049 Munich/Pullach, Germany

and

Ludwig-Maximilians-Universität Munich, Institute of Medical Psychology,
Goethestr. 31, D-80336 Munich, Germany

Albrecht von Müller
Parmenides Foundation, Kirchplatz 1, D-82049 Munich/Pullach, Germany

For other titles published in this series, go to
www.springer.com/series/7816

Britt M. Glatzeder · Vinod Goel
Albrecht von Müller
Editors

Towards a Theory of Thinking

Building Blocks for a Conceptual Framework

Springer

Editors
Britt M. Glatzeder
Parmenides Foundation
Kirchplatz 1
D-82049 Munich/Pullach, Germany
bg@parmenides-foundation.org

Albrecht von Müller
Parmenides Foundation
Kirchplatz 1
D-82049 Munich/Pullach, Germany
avm@parmenides-foundation.org

Vinod Goel
Department of Psychology
York University
Toronto ON, Canada
vgoel@yorku.ca

and

Department of Psychology
University of Hull, UK

ISBN 978-3-642-03128-1 e-ISBN 978-3-642-03129-8
DOI 10.1007/978-3-642-03129-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009932418

Cover design: WMXDesign GmbH, Heidelberg, Germany

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Series Preface

What is Thinking? – Trying to Define an Equally Fascinating and Elusive Phenomenon

Human thinking is probably the most complex phenomenon that evolution has come up with until now. There exists a broad spectrum of definitions, from subsuming almost all processes of cognition to limiting it to language-based, sometimes even only to formalizable reasoning processes. We work with a "medium sized" definition according to which thinking encompasses all operations by which cognitive agents link mental content in order to gain new insights or perspectives. Mental content is, thus, a prerequisite for and the substrate on which thinking operations are executed. The largely unconscious acts of perceptual object stabilization, categorization, emotional evaluation – and retrieving all the above from memory inscriptions – are the processes by which mental content is generated, and are, therefore, seen as prerequisites for thinking operations.

In terms of a differentia specifica, the notion of "thinking" is seen as narrower than the notion of "cognition" and as wider than the notion of "reasoning". Thinking is, thus, seen as a subset of cognition processes; and reasoning processes are seen as a subset of thinking. Besides reasoning, the notion of thinking includes also nonexplicit, intuitive, and associative processes of linking mental content.

According to this definition, thinking is not dependant on language, i.e. also many animals and certainly all mammals show early forms of thinking. The emergence of more complex syntactical structures, however, led to a self-accelerating expansion – for not to say "explosion" – of thinking skills. Syntax boosts the possibility to deal with complex relations and enables the understanding of conceptual hierarchies as well as of self-referential structures. The latter may be directly related to the development of an autobiographic self.

The purpose of thinking can be defined in a twofold way: from a biological point of view, it can be characterized as the most advanced form of assuring homeostasis. From a philosophical point of view, it can be characterized as the crucial means by which the richness of reality unfolds for us. These different descriptions do not constitute a contradiction; they rather articulate the complementary perspectives of asking for the function and or asking for the sense of thinking.

Logic, which has long been seen as the core feature of thinking is an important, but nevertheless rather small part of what thinking really is. It refers only to the coherence of explicitly reviewable linkages made by thinking operations. In contrast, metaphors and analogies constitute a highly content-related way of connecting mental content that is extremely important for thinking, though they often escape a rigorous logical analysis.

The Relevance of the Phenomenon of Thinking

Complex thinking skills are probably the most characteristic feature of humans, and the following four appear to be of particular importance:

- Thinking is the crucial mechanism through which the richness, interrelatedness, and coherence of reality unfold for us. Thinking can be seen as the "crown" of evolutionary sophistication and it is crucial for answering the question "what makes us human".
- Thinking and what we refer to as reality shape themselves mutually. Major breakthroughs in many of today's most fascinating scientific issues (from trying to grasp how consciousness works to bridging the conceptual gap between quantum physics and gravity) require a better understanding of how thinking shapes reality and how reality shapes thinking.
- Ever increasing complexity and a self-accelerating pace of change characterize our modern world. The highly complex, interrelated dynamics of technological, economical, political, and sociocultural developments constitute new challenges that require further advancements in our thinking skills in order to cope with them.
- In an increasingly knowledge-based economy, thinking as the process by which new knowledge is generated will become the main value generation process.

Being aware of the importance of thinking, it is astonishing how little we understand about how complex thinking actually works and how it is implemented in the human brain. The task of the Parmenides Foundation is to enable advanced, interdisciplinary research on this topic.

The Parmenides Foundation and its Research Agenda

The overall purpose of the Parmenides Foundation is to advance our understanding of one of the most fascinating, characteristic and relevant faculties of human beings: complex thinking. The foundation was established in the year 2000 as a non-profit institution for basic research.

The main activity of the foundation is to run the Parmenides Center for the Study of Thinking which has been established in co-operation with the Ludwig Maximilian

University of Munich. The center is organized similar to a Max Planck Institute. It tries to provide optimal conditions for basic research and interdisciplinary cooperation, with minimized bureaucratic distractions and optional teaching activities at the university. The work of the foundation is based on an interdisciplinary core team of approximately 15 scientists at present, a guest fellow program, and an international faculty of about 30 members. The faculty unites outstanding experts from the neurosciences, neuroinformatics, philosophy, cognitive psychology, linguistics, and evolutionary biology.

At present, we focus on the following areas of basic and applied research on thinking. The main topics of basic research are:

- To develop a conceptual framework (or taxonomy) for the understanding of thinking
- To identify and analyse the neural and neurobiological correlates of thinking
- To understand the complementary features of human cognition such as syntactic language and (self)consciousness
- To become able to reconstruct key aspects of complex thinking by modelling
- To learn more about the ontogenesis of complex thinking in childhood
- To learn more about the phylogenesis of complex thinking during evolution
- To study the structural constraints of thinking and their relation to problems in the categorial foundations of science

The main topics of applied research are:

- To develop new approaches and methodologies for supporting the acquisition of thinking skills in early and later childhood
- To develop new approaches and methodologies for supporting the human brain in dealing with tasks of high complexity
- To develop new approaches and methodologies for analyzing and improving the knowledge metabolism of institutions
- To develop new approaches and methodologies for supporting strategy development and decision making in a brain-adequate way
- To develop new approaches and methodologies for the medical reconstruction or restitution of advanced thinking skills

The book series "*On Thinking*" was established to present new insights and findings, as well as ongoing discussions to a wider readership. The volumes are edited by authors from the Parmenides Foundation and Faculty as well as by guest authors and present the progress in this important field for society.

Munich *Ernst Pöppel*
November 2009 *Albrecht von Müller*

Towards a Theory of Thinking

Building Blocks for a Conceptual Framework

.... we do not yet yet know all the basic laws: there is an expanding frontier of ignorance.
Richard Feynman

The mind's characteristic feature, thinking, has been a main theme of philosophical enquiry since the beginning of western philosophy 2,600 years ago. Only recently has it moved out of the philosopher's armchair into the laboratory of scientific research.

Wilhelm Wundt was the first to turn the experimental method of investigation onto the complexities of mental life.[1] He set up the first laboratory for experimental psychology in 1879 in Leipzig. A similar lab was later established in the United States at Cornell University by E.B. Titchener, an American who had studied in Wundt's lab.

At the same time William James dedicated two chapters, "The stream of thought[2]" and "Reasoning," in his seminal book "The Principles of Psychology" (1890) to thinking and thus included it as a topic into the new "science of the mind". James of course, himself both a philosopher and a psychologist, wrote the book before the separation of psychological science and philosophy.

The most active and prolific period in the scientific study of thinking was triggered by a group of philosophers and scientists, who came to be known as the "Würzburg School" of "Denkpsychologie" in the early part of the 20th century. Another important current at this time was the Berlin Institute of Psychology founded by Carl Friedrich Stumpf, which gave birth to Gestalt psychology and extended the Gestalt notion to thinking and reasoning. The main adherents were

[1]Interestingly enough Wundt saw thinking as a collective/social process and sought to understand it within the framework of what he called "Ethnopsychology" (Völkerpsychologie). The social, relational aspect of thinking turned out as one of the more promising approaches to understanding thinking and the human mind today. See part 4, esp. chapter 16 in this book.

[2]James' concept of thought means more than thinking in the narrow sense and refers to conscious mental processes in general. "I use the word thinking for every form of consciousness indiscriminately. If we could say in English 'it thinks,' as we say 'it rains' or 'it blows,' we should be stating the fact most simply and with the minimum of assumption. As we cannot, we must simply say that *thought goes on*." William James, (reprint 1950) The Principles of Psychology, vol. 1, chap. 9 "The Stream of Thought". New York: Dover Publications.

Stumpf's students Wolfgang Köhler, Kurt Koffka, Max Wertheimer, and Kurt Lewin.[3] Two chapters in this volume deal with these important ideas and developments (Chapter 1 by Michael Öllinger and Vinod Goel, and Chap. 4 by Michael Wertheimer).

J.B. Watson and his school of behaviorism overthrew the research program of "Denkpsychologie" and Gestaltheory of thinking by denying the legitimacy of mental concepts, such as thinking and by asserting that the only legitimate and true object of psychological investigation was observable behavior.

It has indeed turned out to be extremely difficult to operationalize thinking as an inner process, and with modern experimental psychology increasingly focussing on operationalization, the concept of thinking has been more and more neglected. In addition, the initial attempt to study the complexities of human thinking as an ongoing mental process as such has been replaced with the separate study of aspects of thinking such as problem solving, concept formation, categorization, or inductive reasoning.

In the last 30 years, the focus of research has gravitated "downwards" to more and more elementary low-level cognitive operations. This approach, typically exemplified by "button pressing experiments", is certainly relevant in its own right. But it is also evident that it is far from studying human thinking in its richness and complexity.

In the present second volume of the Parmenides book series, we seek to tie in with the overarching goal of German "Denkpsychologie" to take thinking seriously as a scientific concept and to go for an integrated study of complex thinking.

While the first volume of the Parmenides book series "On Thinking" is about neural correlates of thinking, the second volume focuses on assembling the building blocks of a conceptual framework that might – after several iterations – develop into a future theory of thinking.

We are, of course, aware that we are still far away from a comprehensive understanding of full-blown human thought, and hence the reference to a "theory of thinking" in the title of this book is not a claim, but a Leitmotiv.

We deem our endeavour to be worthwhile in particular with regard to the present situation in the highly multidisciplinary field of research on higher cognition that resembles what Aristotle characterized as knowing many details – but not understanding the essential phenomenon in its entirety and coherence. For Aristotle, real science begins with striving for the latter. Regarding human thinking, this is a tall order, but at the same time, an unavoidable one.

One of the big challenges for a future theory of thinking is that it calls for intense collaboration between specialists in many fields from molecular and behavioural sciences all the way to the humanities. This volume touches on a

[3]Jean Matter Mandler and George Mandler (1964) Thinking: From Association to Gestalt. New York, London: Wiley & Sons; Robert Sternberg and Edward Smith (eds) (1988) The Psychology of Human Thought. Cambridge: Cambridge University Press; Richard E. Mayer (1991) Thinking, Problem Solving, Cognition. New York: Worth Publishers.

broad range of sources that are potential contributors to a research framework and eventually to a theory of thinking. It brings together an international group of leading scientists coming from the different fields upon which a theory of thinking must build: brain and cognitive sciences; experimental, social, and developmental psychology; evolutionary anthropology and biology; linguistics; neuro-informatics; modeling; and philosophy.

Structure of the Book

A theory of thinking presumably presupposes a common concept of what thinking is. To date, though there are many different concepts of thinking – and neither a generally accepted definition nor an agreement on the exact mechanisms underlying thinking processes. Hence in this volume, we approach our topic by assembling an array of different perspectives on thinking from different scientific fields and explanatory levels – assuming that this strategy allows us to home in and get a better grip on the multifarious phenomenon of thinking.

To organize this "perspectivistic" endeavor – and this volume – we apply a coarse grid that divides the 23 chapters of the book into five main sections. Each chapter covers a pertinent topic of the study of thinking and provides a – major or minor – building block of the emerging conceptual framework. This is of course a selection and by no means would we want to claim it to be complete. But we do think that this compilation exhibits important elements and aspects, which an envisioned framework for a theory of thinking would want/need to integrate.

Part I: Perspectives on Thinking

Within cognitive psychology, thinking has been studied under the headings of problem solving, reasoning, judgment, and decision-making. The first three chapters lay out this landscape and highlight some of the issues.

We start and set the stage with a chapter on problem solving. Problem solving is a field of research with a long tradition and was prominent in the work of psychologists who were the first to develop testable theories in the early part of the twentieth century. To these pioneers in the aforementioned "Denkpsychologie", problem solving was to a large extent thinking per se and this orientation has been shaping research on thinking to this day. Michael Öllinger and Vinod Goel review the most influential theories of problem solving with a revealing focus on the connection between the German Gestalt psychologists and the subsequent further developments within the framework of information processing theory. Particularly Newell's and Simon's Problem Space Hypothesis, which formalizes and builds upon a number of the ideas and findings of the Gestalt psychologists, is detailed. On the basis of a critical assessment of the limits and weaknesses of both approaches, the chapter

closes with presenting an integrative model for insight problem solving which might provide a prototype theory of thinking (understood as problem solving).

In Chap. 2, we turn to reasoning and decision-making. Since psychological studies of reasoning and decision-making started in earnest in the late 1950s, numerous studies have shown that people make systematic errors in their reasoning processes. They often rely on intuitions and gut feelings instead of on more demanding, deliberative reasoning.

Does this force us to conclude that humans are not rational after all? Wim de Neys sees the crucial question for our view of human rationality depend on whether or not people detect that their intuitions conflict with more normative considerations. He reviews recent conflict detection studies that started addressing this issue and suggests that clarifying the efficiency of the conflict detection process and the resulting nature of the heuristic bias is paramount for the development of reasoning and decision-making theories.

Analogy has been another focus of extensive research over the past two decades and is often seen as the very core of human thinking. In Chap. 3, Dedre Gentner and Julie Colhoun argue that the analogical ability to perceive and use purely relational similarity is the major contributor to our species' cognitive uniqueness. While similarity (see Chap. 7) as one of the great forces of mental organization holds across species, only humans experience a sophisticated form of this force, i.e. analogy. The authors present an overview of analogy and describe its component processes. They discuss how these component processes lead to learning and the generation of new knowledge, and review evidence that suggests that greater use of analogy can improve learning.

We then (re)turn to Gestalt theory with a particular interest in the concept of Gestalt, which in our view provides a unique model for a nonsequential, holistic, "constellatory" mode of thinking. In Chap. 4, Michael Wertheimer gives a general overview and an in-depth description of this most basic concept of Gestalt psychology.

The last chapter in Part I offers a philosophical perspective about the relationship between thinking and reality. While the issue has been of major interest to philosophers, it has lost relevance in the laboratories of cognitive science. Albrecht v. Müller argues that the scientific quest for understanding thinking is inseparably linked to the endeavor to understand reality. Informed by contemporary physics, he discusses two complementary aspects of reality and interprets the human faculty of complex thinking as the most advanced evolutionary adaptation to the ambiguous character of physical reality.

Part II: Components of Thinking

Thinking in all its various forms – be it solving a problem, making a decision, or daydreaming – recruits a complex set of cognitive processes that can potentially be applied to a wide range of domains. Some of these constituents of complex thinking are the subject matter of Part II.

We start with categorization (Chap. 6), which is regarded as one of the fundamental abilities of our cognitive system. Categorization is essential for perception and higher cognitive functions. An understanding of how we categorize is central to any theory of thinking. Markus Graf focuses on a fundamental aspect of categorization, viz. to visually recognize and categorize objects. Graf reviews the literature on the background of the hierarchy of transformation groups specified in Felix Klein's "Erlanger Programm", which he proposes as a general framework for the understanding of the recognition of object shapes. The transformational framework proposes that conceptual representations have a similar format as the perceptual input.

The transformational framework plays a role also in the next chapter, which deals with another integral component of human thought: comparison. The process of comparison is crucial in problem solving, judgment, decision-making, categorization, and cognition broadly construed. In turn, the determination of similarities and differences plays a critical role for comparison. Thus, virtually every cognitive process is influenced by implicit or explicit similarity comparisons. As William James put it: "This sense of sameness is the very keel and backbone of our thinking". In Chap. 7, Robert L. Goldstone, Sam Day, and Ji Y. Son review work on comparison and similarity and describe important classes of formal models of these core concepts, viz. geometric, featural, alignment-based, and transformational (see Chap. 6) models.

In Chap. 8, Michael Waldmann deals with causal reasoning, a third component of paramount importance for human thinking. The chapter reviews work that is based on the causal-model approach to causal thinking and learning. It focuses on the contrast between this more recent rational approach and traditional associationist theories, and discusses the causal model theory in light of experimental evidence. (See also causal thinking in animals, Chap. 15 by Josep Call).

To make statements of causality, languages such as English or German commonly use conditionals, i.e. sentences of the form "if…then", but logically conditionals are *not* statements of causality. Conditionals are probably the most important means of expressing our beliefs about how the elements of our world are joined together. We use them to denote causal relations and diagnostic ones, observed regularities, and normative rules, to name just a few. Moreover, conditionals have a prominent role in our reasoning. Chapter 9 by Klaus Oberauer reviews psychological research on reasoning that has recently entered the perennial debate on conditionals in philosophical logic. It evaluates theories of representations and cognitive processes involved in thinking about conditionals, and empirical evidence speaking to two questions: How do people understand conditionals and how do they use them in reasoning? Experiments on people's interpretation of conditionals support the probabilistic view, whereas experiments on reasoning provide evidence favoring the truth-functional view, represented in psychology by the theory of mental models. The author suggests a dual-process account to reconcile the two views.

Memory and perception are preconditions of thinking. Although traditionally conceived as so-called low-level processes, they do not belong to thinking proper, an integrative theory of thinking must take account of these fundamental functions without which our thinking would be devoid of any meaningful content and would not be thinking at all.

Thinking requires varied types of memory. In Chap. 10, Matthias Brand and Hans Markowitsch focus on episodic memory, which ranks highest in Endel Tulving's hierarchy of long-term memory systems and is considered to be a unique feature of humans. Among the characteristic modes of thinking recruiting episodic memory are self-reflection and the ability to mentally travel into past and future. The authors give an overview of studies with brain-damaged patients and investigations employing functional neuroimaging techniques, which provide insights into the neural correlates associating thinking and memory.

Apart from memories, sense perceptions are prime providers of content or material of thinking with vision being considered to play a privileged role. Visual perception has been of great interest to philosophers, psychologist, and neuroscientists alike. A perennial controversy has been going on between two major camps: One argues that perception is a stimulus-driven, bottom-up process, whereas the other contends that it is a constructive, concept-driven process, which relies on top-down processing. In the last years, neurophysiological experiments have approached this problem directly by measuring neural signals in animals as they experience visual percepts. Chapter 11 by Nikos Logothetis[4] reports the results of single-cell recordings of neural activity during binocular rivalry tasks. The findings suggest that there is not one single mechanism or even a single brain area that is responsible for the interesting suppression effects in the context of binocular rivalry, but that neural events operating at distributed networks throughout the visual hierarchy contribute to the overall effect. Moreover the observations seem to dissolve the aforementioned controversy by supporting the idea that part of what we perceive comes through our senses from the "things" around us and another part – which is most likely to be the largest part – comes out of our own mind. (See also Chap. 15 by Josep Call, who provides evidence for top-down processes in apes.)

Part III: Onto- and Phylogenetic Aspects

In the last few decades, more and more researchers investigating human thinking have recognized the value of incorporating developmental and evolutionary neurosciences perspectives in their work. The integration of these perspectives has led to an enriched understanding of human thought processes.

The first chapter of part III presents arguments for the developmental approach, taking modularity as a case in point. For more than thirty years, modularity has been the subject of heated controversies in the cognitive and brain sciences and has shaped the field as well as, some say, impeded progress. Two influential theoretical positions frame the ongoing debate: One claims that the mind/brain is a general-purpose

[4]The chapter is based on a talk presented at the Parmenides faculty meeting 2007. We thank Thomas Filk for the transcription.

problem solver[5]; the other asserts that it is made up of special-purpose modules.[6] The idea that the brain is composed of specialized, independently functioning modules has a long history and dates back to Kant's faculty theory and Franz Josef Gall's phrenology. Jerry Fodor's book *The Modularity of Mind*, in which he provided a precise list of criteria about what constitutes a module, set the stage for the current discussion in cognitive and brain sciences.

The issue of modularity has remained controversial in cognitive and brain sciences for a number of reasons, including: (1) the usefulness of modular structures in engineering; (2) a reemergence of the nature versus nurture debate, where modularity has often been seen as conceptually supportive of a nativist position (e.g. Fodor); (3) the extension of the modularity thesis to higher order thought processes, referred to as "Massive Modularity"; and finally, (4) progress in neuroscience, where new anatomical, imaging, and experimental data have identified a number of brain modules at various levels of granularity.

At the same time, many suggestions have been made in recent years for modifying the Fodorian concept of a module. One big issue in this regard is the claim that innateness and "rigid" cerebral localization are not crucial to modularity. One could very well argue, as Annette Karmiloff-Smith does in Chap. 12, that modules are the result of a gradual process of modularization. She discusses modularity from a developmental perspective and shows how specialization and localization of cognitive and brain function develop constantly across the life-span. The notion that the mind/brain is composed of independent modules may thus hold to some extent for the adult brain, once it has become fully specialized or when it displays acquired domain-specific deficits when focal damage has occurred. However, the extension of this thinking to typically and atypically developing infants in terms of innately specified, intact or impaired modules is, according to Annette Karmiloff-Smith, not warranted.

Research in autism has provided data for the innate modularity versus developmental hypotheses. Adherents of the modularity thesis interpret developmental failures characteristic of autism as a failure of a module devoted to Theory of Mind (ToM). In Chap. 13, Beate Sodian and Susanne Kristen review recent findings on the development of ToM in infancy, which have heated up the debate on modularity and an alleged ToM module. Neuroimaging studies of ToM reasoning in adults provide some support for a specific ToM network. This claim is contested, however, and many relevant studies have yet to be done. The authors conclude that despite nearly 30 years of research effort, there is still no hard evidence for a ToM in nonhuman primates and it seems that ToM is, as widely believed, one of the few uniquely human competences.

In shifting the focus of research from what children know about somebody else's mind to exploring children's awareness of their own cognition, the following chapter

---

[5]Newell, A., and H. Simon (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice Hall; Piaget, J. (1971). Biology and Knowledge. Chicago: University of Chicago Press.

[6]Chomsky, Noam (1980). Rules and Representations. New York: Columbia University; Fodor, J. A. (1983) The Modularity of Mind. Cambridge, MA: MIT Press; Gardner, Howard (1985). Frames of Mind: The Theory of Multiple Intelligences. London: Heinemann.

by Wolfgang Schneider addresses the closely related and equally hot scientific topic of *metacognition.* The paper describes current trends in research on the development of metacognitive competencies, emphasizing the important roles of both procedural and declarative metacognition and reviews major findings on the development of these two components of metacognition. Furthermore, Wolfgang Schneider analyzes the relation of ToM and metacognition and presents data, which support the hypothesis that early ToM competencies can be considered as a precursor of subsequent metamemory. The chapter ends with an outlook on practical applications of metacognition to various educational settings.

One of the most exciting developments has been the move to analyze aspects of animal behaviour that were previously thought to be exclusively human. Wolfgang Köhler, one of the founders of Gestalt Psychology, was famous for his animal studies. By systematic observations he assigned thinking abilities to chimps (like problem solving and insight) that until then were not supposed to occur in animals. In this tradition, research on animal cognition has grown exponentially in the last decade and has established close links with human cognition to jointly explore the mechanisms, the ultimate functions, and the evolution of cognition. Animal cognition has a lot to offer to the study of human thinking and Josep Call, the author of Chap. 15, believes that for some questions about human cognition, animal cognition holds the key answers. However, much of animal cognition is routinely reduced to associations between stimuli and responses and Josep Call argues that this view is too narrow, in particular with regard to apes' causal knowledge about object–object relations. On the basis of the latest experimental studies, he proposes instead that apes distinguish between arbitrary and causal relations between objects. This means that apes not only associate the presence of certain stimuli with certain events but also attribute a causal role between the presence of those objects and certain events. (see Chap. 9 on causal thinking in humans). The chapter closes with some considerations regarding the nature of nonhuman knowledge about the world.

Part IV: Language, Emotion, Culture

In part IV, we adopt a more Vygotskian perspective and zoom out from the focus on the single thinking individual or brain to individuals interacting with other individuals as well as with their environment. The hypothesis that the enlarged brain size of the primates was selected for by social, rather than purely ecological, factors has been strongly influential in studies of primate cognition and behavior over the past two decades. Recent evidence from evolutionary theories, cognitive archeology, paleoanthropology, and embodied cognition theories have reinforced the position. However, traditional cognitive and brain sciences continue to study human cognition by focusing on cognitive and brain processes within single minds/brains.

In Chap. 16, Anne Böckler, Günther Knoblich and Natalie Sebanz consider what can be gained by choosing a new unit of analysis that comprises more than single minds. The authors discuss three kinds of approaches that argue for socializing

cognition by (1) distributing cognitive processes across individuals and the environment, or (2) focusing on evolution and culture as shaping forces, or (3) exploring the functionality of perception-action links. Taken together, the research reviewed in this chapter suggests that respecting the social nature of human cognition will foster a better understanding of individual thinking.

Most higher primates are social by nature. However, the use of language in communication is one of the most conspicuous traits that distinguishes *Homo sapiens* from other species

The origins of language are still the subject of much speculation, but a common theme of many theories – which goes back to Aristotle – is that they are to be traced back to sociality (interactions of many individuals.) The relation between thinking and language has intrigued scholars and writers for centuries and has been a subject of controversy.

Chapter 17 by Per Aage Brandt presents the perspective of Cognitive Semio-Linguistics disclosing two "semiotic bridges" between communication and cognition in our species. Cognitive Semio-Linguistics studies the relations between signs and language, and interprets the semiological and linguistic structures as expressions of, and as causes of, the cognitive activities involved in thinking. "Semiological" refers to the whole of nonlinguistic signs such as symbols, icons, diagrams, traces, or symptoms. These semiological expressions are understood as both more directly connected to the process of thinking and more directly shaped by the structure of the process of thinking. Most often, Per Aage Brandt argues, verbal language does not express thinking directly but interprets the more 'authentic' symbolico-iconic signs of our thinking to render them socially communicable. Brandt views the functioning of the human mind as the 'dialectics' between the linguistic and the semiological bridges. His chapter provides an architectural model of the relations holding between semio-linguistic structures and structures of thinking processes.

Among social species, emotions do not only play a crucial role in regulating social interactions. Emotional states and moods deeply modulate what and how we think.

Chapter 18 by Annette Bolte and Thomas Goschke reviews empirical findings showing that positive and negative affective states are accompanied by qualitatively different information-processing modes. Specifically, positive moods and emotions appear to be associated with a more flexible processing mode as indicated by a broadened scope of attention, activation of weak or unusual associations, and facilitated switching between cognitive sets. The authors interpret these findings within a general theoretical framework according to which different modes of thinking serve complementary or even antagonistic adaptive functions in the planning and control of goal-directed action. In contrast to the widespread view that positive affect has exclusively beneficial consequences such as increased creativity and flexibility, Annette Bolte and Thomas Goschke argue that different emotions and moods and the processing modes associated with them incur complementary costs and benefits. Thus, consistent with recent findings, positive and negative affect have advantages and disadvantages, depending on the processing requirements of the task.

Part IV of our book about the influence of ongoing collective social processes on human brains and minds closes with a chapter by Shihui Han reviewing recent

findings in the relatively novel field of transcultural neuroimaging. Cross-cultural psychological research has provided ample evidence for dissimilar thinking styles in different sociocultural environments. Specifically, people from Western cultures (Europeans and Americans) generally think in an analytic style that is attuned to salient focal objects but less sensitive to contexts, whereas people from East Asian cultures (Chinese, Japanese, Korean) think in a more holistic style that is attuned to background and contextual information. Inspite of the evidence for the diversity and the dependence of human cognition on sociocultural contexts, the question of whether the neural correlates of human cognition are also culture-dependent is rarely considered by neuroscientists who often implicitly assume neurophysiological mechanisms of cognitive processes to be universal. However, recent transcultural neuroimaging studies showed that cultures not only shape multiple-level cognitive processes but also induce variation of neural activity underlying both high- and low-level cognitive processes. These findings help us understand how cognitive processes and the underlying neural mechanisms are modulated by culture and give rise to culture-specific thinking styles.

Part V: Modeling and Neurobiological Approaches

In the fifth and last part of the book, we address the question of how thinking processes may be modeled. This is an important issue in a theory of thinking because modeling approaches can guide the formation of a scientific understanding of thinking by providing a list of constituents that are sufficient for thought processes.

The first in this short suite of chapters presents a strongly "biology-inspired" approach to modeling and envisages natural selection as a model for what goes on in the brain. Ever since Darwin came up with his dangerous idea, philosophers, artists, and economists have been working with Darwinian thought patterns and analogies, and several contemporary theories of mind and brain place the principle of selection at their very center. One of the most recent proposals in this vein comes from Chrisantha Fernando and Eörs Szathmáry, who explore the hypothesis that natural selection takes place in the brain emphasizing the benefits of true replication operations. In Chap. 20, they give a review of the theoretical and experimental evidence for selectionist and competitive dynamics within the brain and propose that to explain certain kinds of productive thinking and behavior, selectionist mechanisms demand extension to encompass the full Darwinian dynamic that arises from introducing replication of neuronal units of selection. They introduce three possible neuronal units of selection, show how they relate to each other, and suggest how these replicators may take part in diverse aspects of cognition such as causal inference, human problem solving, creative thinking, and memory.

It has been stated by many investigators in neuroscience and cognitive science that numerous aspects of human cognition, including high-level cognitive functions involving conscious thought processes, continuously make reference to the internal state of the body, and thus to the fundamental and evolutionarily adaptive systems

of the organism. Chapter 21 by Olaf Sporns and Edgar Körner focus on this feature from the perspective of the emerging field of neurorobotics. Neurorobotics, or embodied artificial intelligence, aims to combine mechanisms and concepts from neuroscience with the design of robotic platforms. They argue that capable neurorobotic systems need to incorporate a flexible and dynamic architecture that supports the self-organization of value and knowledge representation by means of self-referential control. On the basis of a brief review of empirical and theoretical work addressing this area, they outline a set of design principles for a self-organizing and open-ended knowledge architecture, and provide a strategy for its implementation in intelligent systems.

In Chap. 22, Giorgio Innocenti speculates about the hypothesis that in thinking, cortico-cortical connections perform similar associative operations as in perception. The implications of this hypothesis are, he claims, that thinking would be the projections onto the world of cortico-cortical connectivity rather as perception is the projection onto the world of cortico-cortical visual connections. Cortico-cortical connectivity, however, must generate percepts or thoughts compatible with the "real world". Thus, cortico-cortical connections are under a double selective screening performed by evolution and development.

In the very last chapter, Helge Ritter provides an engineering perspective on modeling cognitive processes. He considers methods and requirements that go into the process of model building for the sake of supporting, extending, and boosting human thinking skills. He offers a brief synoptical discussion of some major modeling methodologies and discusses these from a number of complementary dimensions raising issues about how models can aid our thinking, what can be delivered by dynamical systems, how to cope with uncertainty, and how modeling is connected with learning.

Acknowledgements

We thank our distinguished authors and our colleagues at Parmenides for their expertise and cooperation. Special thanks go to Michael Öllinger and Maria Klatte for their scientific and emotional support. Many thanks go also to the team at Springer, Dieter Czeschlik, Anette Lindqvist, and Gnanamani Umamaheswari for supporting our idea of a book series on thinking.

Contents

Contributors

Annette Bolte
Department of Psychology, Technical University of Dresden, Germany
bolte@psychologie.tu-dresden.de

Anne Böckler
Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, A.Bockler@donders.ru.nl

Mathias Brand
University of Bielefeld, Department of Physiological Psychology, 33501 Bielefeld, Germany, m.brand@uni-bielefeld.de

Per Aage Brandt
College of Arts & Sciences, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7068, USA, peraage.brandt@case.edu

Josep Call
Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig Germany, call@eva.mpg.de

Julie Colhoun
Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208, USA

Sam Day
Indiana University, 1101 E 10th St. Bloomington, IN. 47405-7007

Wim De Neys
Experimental Psychology Lab, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium, wim.deneys@psy.kuleuven.be

Chrisantha Fernando
University of Sussex, ctf20@sussex.ac.uk

Dedre Gentner
Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208, USA, gentner@northwestern.edu

Vinod Goel
Department of Psychology, York University, 4700 Keele St., Toronto, ON, Canada M3J 1P3, vgoel@yorku.ca

Robert Goldstone
The Department of Psychological and Brain Sciences, Indiana University, 1101 East Tenth Street, Bloomington, IN 47405, USA
rgoldsto@indiana.edu

Thomas Goschke
Department of Psychology, Technical University of Dresden
goschke@psychologie.tu-dresden.de

Markus Graf
Department Bülthoff, MPI for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany, markus.graf@tuebingen.mpg.de

Shihui Han
Culture and Social Cognitive Neuroscience Laboratory, Department of Psychology, Peking University, 5 Yiheyuan Road, Beijing 100871, China, shan@pku.edu.cn

Giorgio Innocenti
Division of Neuroanatomy and Brain Development, Department of Neuroscience, Karolinska Institute, Retzius väg 8, 171 77 Stockholm, Sweden
giorgio.innocenti@ki.se

Annette Karmiloff-Smith
Developmental Neurocognition Lab, 32 Torrington Square, Birkbeck College, United Kingdom, a.karmiloff-smith@bbk.ac.uk

Günther Knoblich
Donders Institute for Brain, Cognition, and Behavior, Radboud University Nijmegen, 6500 HB Nijmegen, The Netherlands, G.Knoblich@donders.ru.nl

Edgar Körner
Honda Research Institute Europe GmbH, President, Carl-Legien Str. 30, 63073 offenbach/main Germany, Edgar.Koerner@honda-ri.de

Susanne Kristen
Department of Psychology, Ludwig-Maximilians-Universität Munich, Germany

Nikos Logothetis
Department of Logothetis, MPI for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany
nikos.logothetis@tuebingen.mpg.de

Hans Markowitsch
University of Bielefeld, Department of Physiological Psychology, 33501 Bielefeld, Germany, Hans.Markowitsch@uni-bielefeld.de

Klaus Oberauer
Department of Experimental Psychology, University of Bristol, 12a, Priory Road, Bristol, BS8 1TU, UK, K.Oberauer@bris.ac.uk

Michael Öllinger
Parmenides Foundation Kirchplatz 1, D-82049 Munich/Pullach, Germany

Helge Ritter
CITEC – Cognitive Interaction Technology and Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany
helge@techfak.uni-bielefeld.de

Wolfgang Schneider
Department of Educational Psychology, Julius-Maximilians-Universität Wuerzburg, Wittelsbacherplatz 1, Germany
schneider@psychologie.uni-wuerzburg.de

Natalie Sebanz
Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Netherlands

Beate Sodian
Department of Psychology, Ludwig-Maximilians-Universität Munich, Leopoldstr. 13, 80802 München, Germany
sodian@edupsy.uni-muenchen.de

Ji Y. Son
The Department of Psychological and Brain Sciences, Indiana University, 1101 East Tenth Street, Bloomington, IN 47405, USA

Olaf Sporns
Department of Psychology, Indiana University, 1101 E 10th Street Bloomington, IN 47405, USA, osporns@indiana.edu

Eörs Szathmáry
Collegium Budapest, Szentháromság u. 2, H-1014 Budapest, Hungary
szathmary@colbud.hu; Parmenides Foundation, Kirchplatz 1, D-82049 Munich/Pullach, Germany

Albrecht von Müller
Parmenides Foundation, Kirchplatz 1, D-82049 Munich/Pullach, Germany
avm@parmenides-foundation.org

Michael Waldmann
Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany
michael.waldmann@bio.uni-goettingen.de

Michael Wertheimer
Department of Psychology, University of Colorado, Muenzinger Psychology Building, Campus Box 345, Boulder, CO 80309, USA
Michael.Wertheimer@colorado.edu

Part I
Perspectives on Thinking
Building Blocks for a Conceptual Framework

Problem Solving

Michael Öllinger and Vinod Goel

There is no problem so big it can't be run away from.

– Charles Schultz

Abstract Problem solving and thinking are inseparably linked together. We propose that a theory of thinking has to consider and incorporate the notion of problem solving. In this chapter, we review the most important accounts of problem solving and hope to convince the reader that problem solving may provide an ideal framework for developing a theory of thinking.

We start with a broad summary on the Gestaltist perspective. The Gestaltists per se understood thinking as problem solving. They invented a large body of theoretical concepts and ingenious tasks that until now influence cognitive psychology in general and unexpectedly affects the development of the information processing account also. However, this influence becomes less and less explicit and is not appropriately recognized. We hope to stress this connection and bring it back to the readers' minds. Nevertheless, the Gestaltist approach has its weaknesses and methodological flaws, which will be dealt with in this chapter.

A large section is dedicated to the information processing account that still dominates the problem solving literature as a clear and proper account for describing and defining human problem solving. We elaborate on the differentiation between well and ill-defined problems and provide several foundations and models derived from this account. Nevertheless, the information processing account has its limits and we conclude with some extensions of the classical account and provide an integrative model for insight problem solving.

M. Öllinger (✉)
Parmenides Center for the Study of Thinking, Munich, Germany
e-mail: michael.oellinger@parmenides-foundation.org

V. Goel
Department of Psychology, York University, Toronto, Canada
Department of Psychology, University of Hull, UK
e-mail: vgoel@yorku.ca

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_1, © Springer-Verlag Berlin Heidelberg 2010

1 Introduction

In this chapter, we will focus on the study of thought processes through the study of problem-solving. Problem solving can be understood as the bridging of the gap between an initial state of affairs and a desired state where no predetermined operator or strategy is known to the individual. For example, consider the following task: $3 \times 4 = ?$ This task does not constitute a problem for most adults. They can automatically produce the result from memory. For a 7-year-old child, on the other hand, who is learning to multiply, it is a problem. The child has to consciously apply rules and procedures to bridge the gap between the initial problem state and a solution state. For a 2-year-old the situation is not recognizable as a problem because the child lacks the knowledge and semantics to understand what he/she is being asked to do.

In the following sections we will review psychological theories on problem solving, beginning with the work of the Gestaltist psychologists, and the subsequent development within the framework of information processing theory, and then point out two outstanding challenges that the information processing framework needs to confront and overcome.

2 The Gestaltist Perspective

At the beginning of the last century, the Gestaltist approach (Wertheimer 1912, 1925, 1959; Koffka 1935; Katona 1940; Duncker 1945; Köhler 1947) emerged as a countermovement to the dominant learning theory of Behaviorism. For the Gestaltist, thinking was not a reproductive recombination of learned associations but the meaningful effort to understand the fundamental nature and affordances of the given problem situation and the desired goal as a whole. They assumed that thinking obeyed similar basic principles (Gestalt laws) as perception. The Gestaltist idea was that, as in the flipping of the Necker Cube, there are also major transitions during the process of problem solving characterized by restructuring the given information in new and nonobvious ways. Restructuring reveals the fundamental structure of the problem. Problem solving was viewed as a process of transforming a disturbed Gestalt into a good Gestalt ("gute Gestalt"). It is a goal directed behavior that clears out existing barriers in the service of gaining a desired end (for an overview see Ash 1998; Öllinger and Knoblich 2009).

Between 1914 and 1917 Wolfgang Köhler investigated chimpanzees on Tenerife island. He addressed the question of whether chimps are able to solve problems in an intelligent way. He hoped to find evidence against the Behaviorist dictum that animals solve problems by pure trial and error (Thorndike 1911; Köhler 1921, 1925). He claimed that intelligent behavior can be observed when the obvious way to the goal is blocked by a barrier. That is, intelligence is used to elude existing barriers in new and unfamiliar situations. He created situations in which his apes had to solve problems. Sultan, the star pupil, was asked to get a banana that was out of reach. There were two sticks lying around in the compound. After a few minutes

Sultan purposefully joined the sticks together and successfully fished for the banana. For Köhler these findings provided evidence that some animals were able to solve problems not simply by blind and mindless trial and error attempts, but by insight into the affordances of the given situation.

Max Wertheimer the most famous and influential Gestaltist was particularly interested in the sudden moment of restructuring that accompanied insight in a given problem. Wertheimer contrasted *productive thinking* (Wertheimer 1959) with *reproductive thinking (Thorndike* 1911*)*. He was certain that productive thinking is superior to reproductive thinking, because it is characterized by gaining deep insight into the relations of the given problem constituents and their role in the given task, and the resulting solution. Wertheimer worked on a general psychological theory of problem solving that can be applied to various phenomena, ranging from low-level perceptual phenomena, to solving problems like crypt arithmetic, to explaining great scientific inventions, to problems in the social domain (Wertheimer 1959). Restructuring was the basic mechanism for resolving problems across a wide range of domains.

The Gestaltists demonstrated what they meant by "restructuring*"* in a series of elegant examples (Wertheimer 1925, 1959). Figure 1a depicts a typical example. The task is to determine the area of the isosceles triangle, given the length of the side s and the angle at the apex (90°). At first glance people might try to determine the two segments *g* and *h* in order to apply the triangle formula ½ *g* × *h* for the area. This is a laborious approach. However, rotating the triangle reveals that the triangle can be understood as one half of a square with the diagonal *g* and the side with the length s (Fig. 1b). Now, the area can be determined by the simple formula *s* × *s* / 2.

Restructuring the given situation requires a problem solver to overcome the reproductive tendency to compute the triangle area in the usual way and see it in a new way as part of a larger good Gestalt.

Another problem that Wertheimer analyzed was the famous enumeration problem solved by the young Gauss. The task was to add as quickly as possible the sum

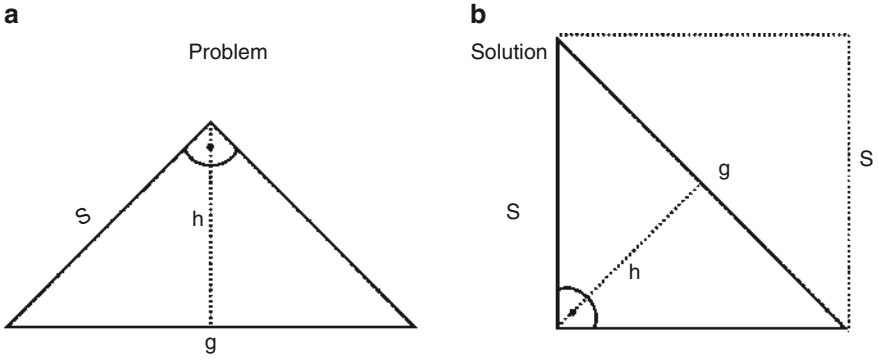

Fig. 1 Wertheimer's Triangle problem (**a**). The task was to determine the area of the given triangle. Insightful solution of the problem (**b**)

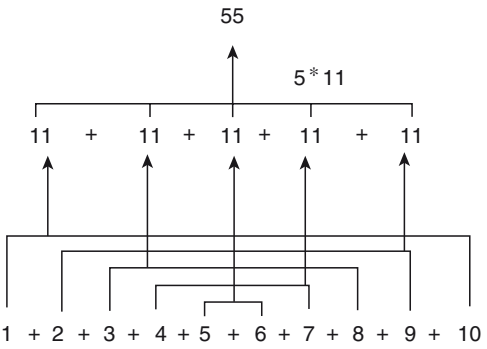

Fig. 2 Gauss' Enumeration Problem. Determine the sum of the given series. The trick is to re-cluster the problem elements and multiply the number of clusters. In this case the value of a cluster is 11 and there are 5 of those clusters

of 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10. Anecdotally, it is said that almost immediately Gauss proclaimed "Here it is!", and he showed his disbelieving teacher the correct solution (see Fig. 2). The reproductive solution requires stupid blind successive addition of consecutive numbers. However, Gauss found a productive solution in terms of a general principle of arithmetic progression ($(n + 1) \times (n/2)$; n = length of a series). Of course, the longer the series the more effective is the productive approach. The reader is invited to try both approaches on a series from 1 to 100.

Probably the most important Gestaltist work on problem solving was reported by Karl Duncker in his Monograph: On Problem-Solving (Duncker 1945). Duncker extended the basic principle of restructuring by a general framework that views problem solving as a stepwise process situated in a problem space which people navigate by means of strategies or heuristics. Duncker anticipated some concepts that later became the fundamentals of Newell and Simon's Problem Space Hypothesis (Newell and Simon 1972).

Duncker also introduced a number of classical problems, like the radiation problem and the candle problem, into the literature. The radiation problem asked the following question:

> Given a human being with an inoperable stomach tumor, and rays which destroy organic tissue at sufficient intensity, by what procedure can one free him of the tumor by these rays and at the same time avoid destroying the healthy tissue which surrounds it? (Duncker 1945, p. 1–2).

This has proved to be a fairly difficult problem. The solution is to use more than one laser of weak intensity and arrange them in a way that their rays exactly meet right in the heart of the tumor. The superimposed radiation destroys the tumor and does no harm to the surrounding tissue. In Duncker's studies, participants were asked to "think aloud" or verbalize thoughts and ideas as they are attended to, while solving the problem. This technique has become an important methodical instrument, within information processing theory, for mapping intermittent steps in thinking processes onto cognitive models (Schooler and Engstler-Schooler 1990; Ericsson and Simon 1993; Schooler et al. 1993, Goel and Pirolli 1992).

Duncker analyzed the thinking-aloud protocols and systematically developed graphs, such as in Fig. 3.

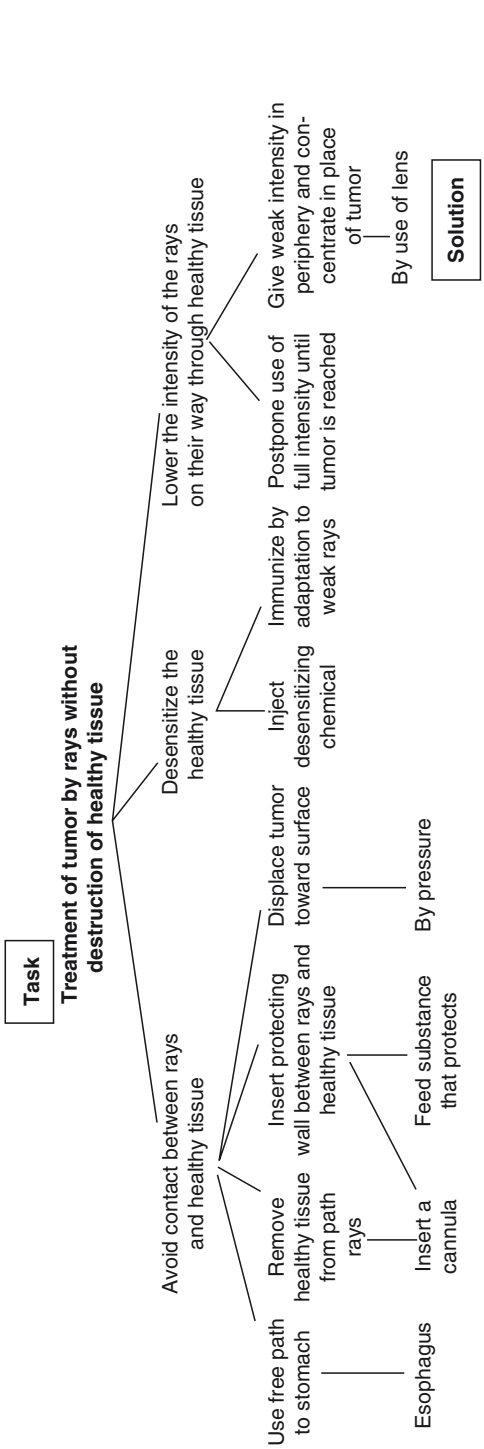

Fig. 3 Incorrect solution attempts and the correct solution for Duncker's radiation problem

This graph classifies different kinds of solution attempts for the radiation problem according to their "functional value". The functional value of a solution "is exactly what is called the sense, the principle or the point of the solution" (Duncker 1945, p. 4). For the radiation problem Duncker identified three basic methods participants used to solve the problem, but eventually only the "Lower the intensity of rays" path led to a correct solution.

The graph illustrates a hierarchical problem solving structure. The top of the graph represents the initial state and constraints given in the task, followed by potential solutions for meeting task requirements. The process of traversing the graph is accompanied by crosschecking for the suitability of steps and their relationship to the solution. Duncker found out that people search in both directions, that is, either from the initial state to the goal state or from a potential goal state back to the initial state (Holyoak 1995). This kind of task analysis is fairly similar to the more recent concepts of problem space and search within information processing theory (Newell and Simon 1972) that still dominates the current problem solving research (see below).

The Gestaltists were also interested in how prior knowledge influences the solution of problems. They assumed that prior knowledge can hamper productive thinking. In his famous Candle Problem, Duncker asked participants to find a way to create a ledge on the wall to place a candle. Matches, a candle, and a box of thumbtacks were placed on a table top in front of subjects. The solution is to remove the thumbtacks from the box, and use the box as a ledge and fasten it with the tacks onto the wall, and rest the candle on the box/ledge. The problem is quite difficult. Only a small percentage of people find the correct solution. Duncker explained the problem by assuming that problem solvers *fixate* on the container function of the box. That is, the experience with boxes as containers to put things in was detrimental to seeing it in another way. In a second experiment he placed the thumbtacks either inside the box or beside the box, thus emphasizing or deemphasizing the container function of the box. He found that in the case where the thumbtacks were placed beside the box (deemphasizing its container function) the probability of using the box as a platform was increased.

Further experimental studies provided support for Duncker's assumptions (Maier 1931; Birch and Rabinowitz 1951). Luchins (1942) showed that functional fixedness did not only appear when using objects in an uncommon way. He demonstrated also that the repeated application of the same solution procedure can result in a *mental set* that prevents people from applying alternative and more efficient solution strategies.

Luchins examined mental set by using water jug problems (Luchins 1942; Luchins and Luchins 1959; Luchins and Luchins 1994; Lovett and Anderson 1996; Öllinger et al. 2008). For example, given three jugs A, B, and C, with volumes of 21, 127, and 3 units, respectively, the goal is to end with an amount of 100 units. The solution is to pour water into B (127), then use the water in B to fill C twice, leaving 121 units in B. Now, pour 21 units from B into A.

Luchins created a set of problems that could be solved by the same solution procedure (B – 2 × C – A). After participants learned the solution they were confronted with a test problem that either could be solved with the previously learned

solution procedure or with a simpler procedure. For example, given the volumes 23, 49, and 3 in jugs A, B, and C, with the goal of attaining 20 units, the simple alternative is to pour once water from A into C, and 20 units are left in A. Luchins' experiments showed that participants who had learned a solution procedure based on previous problems, continued to use the same procedure, even though a simpler procedure would suffice. A control group that only solved the test problems applied the easier procedure.

Functional fixedness and mental set became the key concepts of the Gestaltists to explain why productive thinking is often so difficult and why blind and stupid drill in school is detrimental for creative and insightful problem solving. Although the Gestaltists provided a number of valuable ideas and concepts (Novick and Bassok 2005) their theoretical and empirical contributions, and language were sometimes vague, unclear, phenomenological, and hard to formalize. In the next section we will introduce the information processing theory and the Problem Space Hypothesis (Newell and Simon 1972), which formalizes and builds upon a number of the ideas and insights of the Gestalt psychologists.

3 Information Processing Theory and the Problem Space Hypothesis

After the Gestaltist phase there was a short moratorium in problem solving research. In 1957 Newell and Simon presented their General Problem Solver (GPS) at the Dartmouth Conference (Newell et al. 1958). This meeting ushered in a new framework, accompanied by great optimism, that sooner or later thinking processes may be described and understood as computational processes (Ernst and Newell 1969). In the following sections we will review some of the key ideas and contributions of this research program, and then point out some limitations and challenges that it faces.

Information processing theory accounts of human problem-solving appeal to three main notions (see Fig. 4): (a) an information processing system, (b) the task

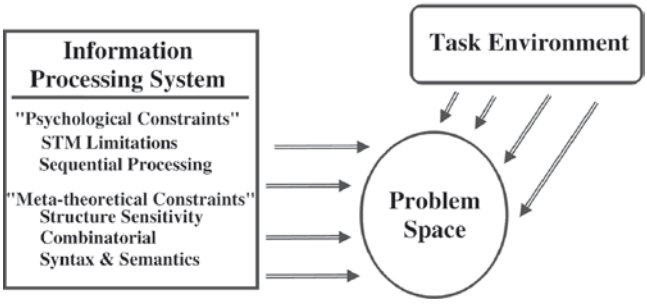

Fig. 4 The problem space is a computational construct shaped by the constraints imposed by the structures of the information processing system and the task environment. Reproduced from Goel (1995)

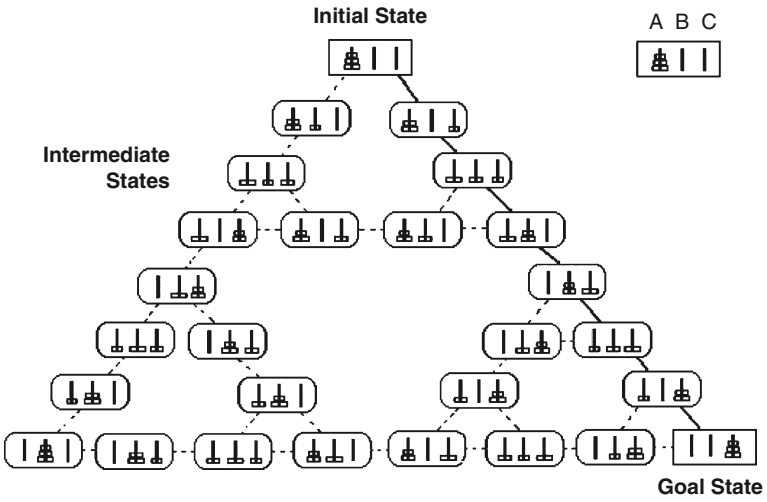

Fig. 5 The Tower of Hanoi puzzle consists of three pegs and several disks of varying size. Given a start state, in which the disks are stacked on one or more pegs, the task is to reach a goal state in which the disks are stacked in descending order on a specified peg. There are three constraints on the transformation of the start state into the goal state. (1) Only one disk may be moved at a time. (2) Any disk not being currently moved must remain on the pegs. (3) A larger disk may not be placed on a smaller disk. This is an example with three disks. The initial state, goal state and intermediate states are indicated. The optimal solution is the right-most sequence of consecutive moves

environment, and (c) the problem space. An Information Processing System (IPS) is a physical symbol manipulation system with memory stores (short term, long term, external), a processor, sensory receptors, and motor effecters. It brings to bear two sets of constraints. The psychological constraints consist of temporal and spatial limitations on working memory and sequential processing. The meta-theoretical constraints require that the information processing system be a computational system with combinatorial syntax and semantics, and structure sensitivity to process (see Fodor and Pylyshyn 1988). Task environments consist of (a) the goal, (b) the problem, and (c) other relevant external factors (Newell and Simon 1972). The problem space is a computational space shaped by the interaction of the constraints inherent in the information processing system and the task environment. It is defined by a state space, operators, evaluation functions, and search strategies.

We can illustrate each of the components of the problem space by reference to the well known Tower of Hanoi problem, depicted in Fig. 5. The state space consists of an initial state, a goal state, and intermediate states. The initial state (in Fig. 5) consists of three disks stacked on the left most peg of the puzzle. The goal state consists of the three disks stacked on the rightmost peg of this example. The initial state is transformed into the goal state by the application of a series of operators/transformation functions, resulting in intermediate states. In the case of the Tower of Hanoi the operator might consist of moving a disk from one peg to another peg, respecting the constraints that only one disk may be moved at a time,

that any disk not being moved must remain on the peg, and that a larger disk may not be placed on a smaller disk. The application of operators to the initial state generates intermediate states.

Evaluation functions determine whether the solution path is moving one closer to or further away from the goal state. These functions may be used to select between different paths through the state space. For example, in the above Tower of Hanoi problem, one can move the small disk onto the middle peg or onto the rightmost peg. Most subjects will reason that if they move the disk to the rightmost peg, it will conflict with the goal state, because the largest peg needs to be placed on the rightmost peg first. Therefore, they will typically move the smallest disk to the middle peg. This actually turns out to be a mistake because it leads to some backward moves or the stacking of the disks on the middle (incorrect) peg.

The application of operators is guided by control strategies. The built-in strategies for searching this problem space include such content free universal methods as Means-Ends Analysis, Breath-First Search, Depth-First Search, etc. Means-ends analysis provides an important and elegant example of how a human like goal-directed behavior can be implemented in a computational system (Newell 1990; Anderson and Lebiere 1998). The application of the means-ends analysis follows three consecutive steps. First, the algorithm determines the distance between the current state and the goal state. Second, it tests whether there is an available operator that reduces the distance. If this is not the case a subgoal is created and pushed into a goal stack. Next, the procedure jumps back to the first step. The means-ends analysis creates subgoals until an operator is available that can be applied, third step. Now, the next stored subgoal is processed. The algorithm terminates when all subgoals are executed.

The strategy can be illustrated by application to the Tower of Hanoi puzzle in Fig. 5. The problem requires the large disk to be transferred to peg C by applying the move-the-largest-disk operator. This is only possible if the medium disk is moved (move-medium-disk operator; subgoal). To do so the smallest disk has to be cleared out of the way (move-small-disk operator, sub-sub-goal). This operator is available and the smallest disk can be moved to peg C. Now, it is possible to go back to the previous subgoal and apply the no longer blocked move-medium-disk operator to peg B. The small disk at peg C blocks the application of the move-large-disk operator therefore the move-small-disk operator is again applied and moves the smallest disk to peg B. Finally, the large disk is moved to peg C etc.

However the universal applicability of these formal methods comes at the cost of enormous computational resources. Given that the cognitive agent is a time and memory bound serial processor it would often not be able to respond in real time, if it had to rely on formal, context independent processes. So the first line of defense for such a system is the deployment of task-specific knowledge to circumvent formal search procedures.

A heuristic strategy for solving the Tower of Hanoi problem might be the following (Simon 1975): (1) on odd numbered moves, move the smallest disk; (2) on even numbered moves, move the next smallest exposed disk; (3) if the total number

of disks is odd, move the smallest disk from the source peg to the target peg, to the other peg, to the source peg, etc. The strategy requires the concepts of odd-even moves and cycling a disk through the pegs in a particular order. The perceptual tests are quite simple, consisting of location of smallest and next smallest exposed disks and differentiating between target source and other pegs. The strategy also makes few computational demands, requiring only the retention of move parity in short-term memory and is therefore very easy to implement. However, it is an "unreasoned" heuristic strategy. It just happens to work for this problem. Gigerenzer and colleagues have undertaken extensive investigations into the nature and role of heuristics in human problem solving (Gigerenzer and Hug 1992; Gigerenzer and Todd, 2001a, 2001b).

The great achievement of the Information Processing Theory was the emphasis on detailed task analyzes, and the clear computational characterization of the problem space. It has resulted in new insights into the nature of certain types of problem-solving. However, 40 years after the onset of the program, the scope of the framework seems limited to a narrow range of problems. It is proving difficult, perhaps impossible, to encompass certain critical types of real-world problem solving within this framework. We've reviewed to search problem types below: Ill-structured problems and insight problems.

4 Challenge: Well-Structured Versus Ill-Structured Problems

The ill-structured/well-structured distinction originates with Reitman (1964). Reitman classified problems based on the distribution of information within the three components (start state, goal state, and the transformation function) of a problem vector. Problems where the information content of each of the vector components is absent or incomplete are said to be ill-structured. To the extent that the information is completely specified, the problem is well-structured.

A mundane example of an ill-structured problem is provided by the task of planning a meal for a guest. The start state is the current state of affairs. While some of the salient facts are apparent, it is not clear that all the relevant aspects can be immediately specified or determined (e.g., how hungry will they be?; how much time and effort do I want to expend?; etc.). The goal state, while clear in the broadest sense (i.e., have a successful meal), cannot be fully articulated (e.g., how much do I care about impressing the guest?; should there be 3 or 4 courses?; would salmon be appropriate?; would they prefer a barbecue or an indoor meal?; etc.). And finally, the transformation function is also incompletely specified (e.g., should I have the meal catered, prepare it myself, or ask everyone to bring a dish?; if I prepare it, should I use fresh or frozen salmon? etc.).

Well-structured problems on the other hand, are characterized by the presence of information in each of the components of the problem vector. The Tower of Hanoi (Fig. 5) provides a relevant example. The start state is completely specified (e.g., the disks are stacked in descending order on peg A). There is a clearly defined test

for the goal state (e.g., stack the disks in descending order on peg C). The transformation function is restricted to moving disks within the following constraints: (1) Only one disk may be moved at a time. (2) Any disk not being currently moved must remain on a peg. (3) A larger disk may not be placed on a smaller disk.

Reitman's original characterization has been extended along a number of dimensions by (Goel 1995). One very important, but little noted, difference has to do with the nature of the constraints in the two cases. In the Tower of Hanoi, as in all puzzles and games, the constraints are logical or constitutive of the task. That is, if one violates a constraint or rule, one is simply not playing the game. For example, if I place a bigger disk on a smaller disk I am simply not doing the Tower of Hanoi task.

However, the constraints we encounter in real-world situations are of a very different character. Some of these constraints are nomological; many of them are social, political, economic, cultural, etc. We will encompass the latter category under the predicate "intentional". In fact one can view social, cultural, religious norms (e.g., Thou shalt not commit adultery! Thou shalt not lie!) as attempts to provide structure to our lives. However, as part of the educational processes, most of us quickly learn that these constraints are not definitional or constitutive of the task. On the contrary, they are negotiable/breakable, depending on circumstances (e.g., maybe it is ok if I don't get caught).

It is also the case that in most ill-structured situations, there are no right or wrong answers, though there are certainly better and worse answers (Rittel and Webber 1974). In the above dinner example, if our dinner guest eats what we serve, did we reach the correct goal state? This seems like an odd question. There will always be better and worse possibilities than any given outcome.

In well-structured problems there are right and wrong answers, and clear ways of recognizing when they have been reached. So if I succeed in stacking my disks in descending order on peg 3 in the Tower of Hanoi task, that is the one and only possible correct answer.

All problems require registration and decomposition, or at least individuation. There are differences with respect to the lines of decomposition/individuation and the interconnectivity of components. Well-structured problems have a predetermined structure, which is either explicitly given with the problem, or is implied by the logical structure of the problem. (So, for example, on a standard interpretation of the game of chess, each player starts with 16 game pieces. One does not have the option of claiming that the conjunction of one of the "rooks" and "knights" constitutes a game piece.).

In ill-structured problems, on the other hand, lines of decomposition/individuation are determined by the subject, taking into consideration the physical structure of the world, social and cultural practices, and personal preference.

In terms of the interconnectivity of parts, one finds logical interconnections in well-structured problems (e.g., in cryptarithmetic there is always the possibility that any row will sum to greater than 9 and affect the next row). Thus the subject has no choice or selectivity in attending to interconnections. Interconnections in ill-structured problems are contingent and one has considerable latitude in determining which ones to attend to and which ones to ignore.

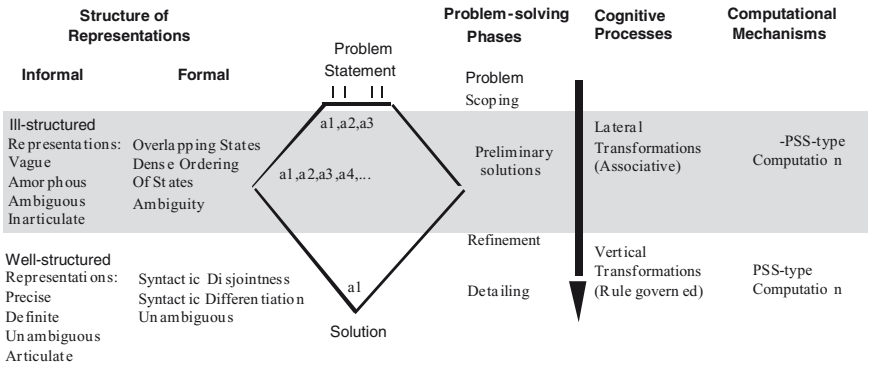

Fig. 6 Aspects of real-world problem-solving. Reproduced from Goel (1995)

Simon (1973) has famously argued that the distinction between ill-structured and well-structured problems is ill-conceived. The so-called "Ill-structured problems" are simply structured by adding information from our background knowledge and external sources and then one can specify the problem space and search for a solution. Subsequent research does not bear out this claim.

Goel (1995) views ill-structured problem-solving as typically involving the following four phases: *problem scoping*, *preliminary solutions*, *refinement*, and *detailing* of solutions. Each phase differs with respect to the type of information dealt with, the degree of commitment to generated ideas, the level of detail attended to, the number and types of transformations engaged in, the mental representations needed to support the different types of information and transformations, and the corresponding computational mechanism (Goel 1995). As one progresses from the preliminary phases to the detailing phases, the problem becomes more structured. This is depicted in Fig. 6.

Preliminary solution generation is a classical case of creative, ill-structured problem solving. It is a phase of "cognitive way-finding", a phase of concept construction, where a few kernel ideas are generated and explored through transformations. This generation and exploration of ideas/concepts is facilitated by the abstract nature of information being considered, a low degree of commitment to generated ideas, the coarseness of detail, and a large number of *lateral transformations*. A lateral transformation is one where movement is from one idea to a slightly different idea rather than a more detailed version of the same idea. Lateral transformations are necessary for the *widening* of the problem space and the exploration and development of kernel ideas. The rules underlying lateral transformations cannot be articulated (Goel 1995).

The *refinement* and *detailing* phases are more constrained and structured. They are phases where preconstructed concepts are manipulated. Commitments are made to a particular solution and propagated through the problem space. They are characterized by the concrete nature of information being considered, a high

degree of commitment to generated ideas, attention to detail, and a large number of *vertical transformations*. A vertical transformation is one where movement is from one idea to a more detailed version of the same idea. It results in a *deepening* of the problem space. The rules underlying vertical transformations can often be articulated (Goel 1995).

Goel (1995) has argued that the ability to engage in lateral transformations is underwritten by a mechanism that supports ill-structured mental representations and computation. Ill-structured representations are imprecise, ambiguous, fluid, indeterminate, vague, etc. The ability to engage in vertical transformations is underwritten by a mechanism that supports well-structured mental representations and computation. Well-structured representations are precise, distinct, determinate, and unambiguous.

Furthermore, there is a computational dissociation between these two mechanisms (Giunti 1997; Goel 1995). Laboratory problems emphasize well-structured mental representations while real-world problems require both ill- and well-structured mental representations. Ill-structured and well-structured representations differ with respect to modes of inference and computational mechanisms. It has also been suggested that *there is an anatomical dissociation corresponding to the computational dissociation* (Goel 2005). If this analysis is correct, it severely limits the scope and relevance of the Newell and Simon framework to our understanding of human problem solving.

5 Challenge: Insight Problem Solving

Insight problems reveal further limitations of the classical information processing theory. The enigma of insight problems is that they mostly have a fairly small problem space and sometimes can be solved by only one single move; however they can be extremely difficult. Insight problems present a challenge for classical Information Processing Theory.

5.1 Definition of Insight

There is no general agreement on the characterization of insight problems. There are phenomenological, task, and process definitions (Knoblich and Öllinger 2006; Öllinger and Knoblich 2009). Two theoretical accounts try to extend the application of information processing theory to insight problems. The first account (Kaplan and Simon 1990) assumes that insight problems are nothing special. A second "process definition" account states that insight problems are characterized by the requirement of a representational change – in Gestalt terms a moment of restructuring, often accompanied by a kind of Aha! experience (Bowden et al. 2005).

5.2 *Nothing Special Account*

Insight problems are difficult because of an ill-defined initial problem representation that integrates far too much information. A much too large problem space is established and therefore an exhaustive search is impossible. A further related point is that in most cases, problem solvers do not have the appropriate heuristics for constraining the search space – that is the transformation function sensu Reitman is not specified. Therefore, insight problems are so difficult because they require the problem solvers to find an appropriate heuristics within an ill-structured problem representation (see also Chronicle et al. 2004; MacGregor et al. 2001; Ormerod, MacGregor, and Chronicle 2002).

> … noticing invariants is a widely applicable rule of thumb for searching in ill-defined domains, [but] there can be no guarantee that those noticed will be the critical ones for the particular problem. Nevertheless, the constraints offered by the notice-invariant heuristic are a vast improvement over blind trial and error search. Kaplan and Simon (1990) p. 404

Kaplan and Simon (1990) investigated their assumptions with the mutilated checkerboard problem (Wickelgren 1974). The task is to determine whether the remaining 62 squares (Fig. 7) can be covered with 31 dominos, or to show that this is impossible. The task was extremely difficult, only a few persons were able to correctly solve it. The answer is, it is impossible to cover the mutilated board. The reason is that two white squares have been removed. And a domino can only cover to adjacent squares of different colors. The problem became significantly easier by providing a version that explicitly emphasizes this aspect (Fig. 7; Bread-and-Butter version).

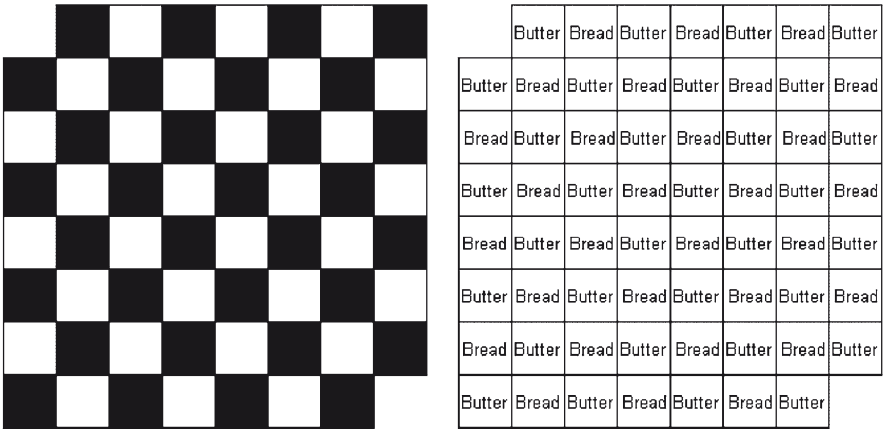

Fig. 7 Wickelgreen's mutilated checkerboard problem. On the left side the original problem, on the right side the Bread-and-Butter version that facilitates finding the solution

5.3 *Representational Change*

The second account (Ohlsson 1990; Knoblich et al. 1999) holds that people apply self-imposed and unconscious constraints on the problem representation. Such an over-constraint problem representation dramatically hampers the solution of the problem. The key mechanism for gaining insight is a representational change. A representational change modifies either the specification of the initial state (chunk decomposition), or the representation of the goal (constraint relaxation, see Knoblich et al. 1999; Öllinger et al. 2008).

During a successful problem solving process three different phases are crossed, namely before, within and after an *impasse*. Before an impasse, problem solving is driven by prior knowledge that suggests that the problem can be solved as usual. For insight problems those attempts normally lead into an impasse. An impasse is defined as a state of mind where problem solving attempts cease and the impression arises that the problem is unsolvable. A representational change either of the given problem situation or the goal representation is necessary. A new and more general representation is established. After an impasse the common strategies (e.g., means-ends analysis) are applied onto the new representation (Öllinger and Knoblich 2009). This assumption fits quite well to Wallas' (1926) four tier model of scientific problem solving. In the *preparation phase* people gather information and make first solution attempts. After a number of failed attempts the problem is put aside and other things come into focus. This phase is called *incubation*. After a while, *illumination* occurs. The key to the solution is found accompanied by insight and an Aha! experience. Finally, in the *verification* phase people verify whether the found solution works.

A prototypical example for the necessity of a change of an over-constraint goal representation is provided by Katona's triangle problem (Katona 1940). The task is to arrange six given matchsticks in a way that four equilateral triangles result. Most participants try to solve the problem in a 2D fashion (see Fig. 8a), but in 2D the problem is unsolvable. The solution requires overcoming the 2D constraint and searching for a solution in 3D.

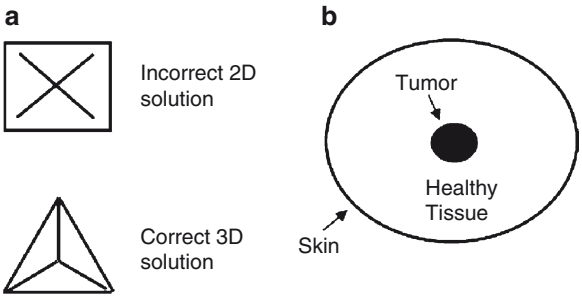

Fig. 8 (**a**) Solution and incorrect solution attempt for Katona's triangle problem. Note: The incorrect solution did indeed produce four equal triangles, but not equilateral triangles. (**b**) Duncker's radiation problem. Sketch used in an eye-movement study by Grant and Spivey (2003)

To investigate the dynamics and preconditions of representational change Grant and Spivey (2003) introduced a very elegant method. In two successive experiments they asked participants to solve Duncker's radiation problem (as introduced in the Gestalt section). Additionally, participants received a sketch (see Fig. 8b) of the problem. In the first experiment participants solved the problem and in parallel their eye movements were recorded. Analyzing the eye movement patterns revealed that solvers looked significantly longer at the skin than nonsolvers did.

The authors concluded that fixating at the tumor – the site where the desired effect should occur – constrains the problem representation and prevent a solution. If so, then, it should be possible to increase people's solution rate by guiding their attention towards the area around the skin. In the second experiment they introduced three conditions. In the first condition the surrounding skin slightly flickered. In the second condition the tumor flickered and finally, in the control condition participants got the static sketch. The result showed that guiding attention to the skin strongly increased the solution rate. This might provide evidence that some insight problems are difficult due to an over-constrained initial problem representation.

5.4 An Integrative Perspective

Currently, researchers are trying to develop more integrative models of insight problems that extend the classical Information Processing Account. There is a long standing and heated debate about the Nine-Dot problem (Maier 1930; Scheerer 1963; Weisberg and Alba 1981; MacGregor et al. 2001; Kershaw and Ohlsson 2004). The task is to connect nine dots with four connected straight lines without lifting the pen or retracing a line (see Fig. 9a). The problem proved extremely difficult. The solution rates usually lay between 0 and 10%.

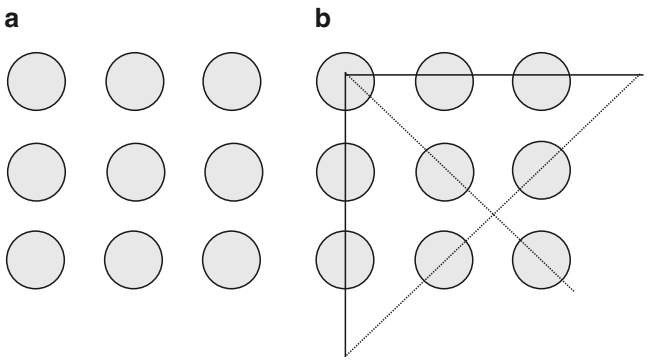

Fig. 9 Nine-Dot problem (**a**) with solution (**b**)

The Gestaltist's standard explanation was that fixation on the 3 × 3 dot matrix that forms a virtual square prevents moves outside the square's boundaries (Fig. 9b). The fixation is the result of perceptual Gestalt laws like figural integrity and figure ground perception (Maier 1931; Ohlsson 1990; Scheerer 1963). Weisberg and Alba (1981) questioned this interpretation. They deduced the following from the fixation assumption: if fixation is the main source of problem difficulty then providing a line that goes beyond the barriers of the virtual square should dramatically increase the solution rate. In their experiments they did exactly this manipulation and found no facilitation. This null-effect was often taken as evidence that Gestalt principles did not play a major role for the solution of the Nine-Dot problem.

However, Kershaw and Ohlsson (2004) demonstrated on the basis of several experimental manipulations that no single source is responsible for the poor solution rate, the interplay of at least three different factors are required. They identified perceptual (Gestalt laws), conceptual (over-constraint goal representation), and process factors (look-ahead, working memory demands) whose interplay impede the solution.

We conclude that understanding the dynamic interplay between problem and goal representation, contextual factors and the process factors, like working memory and executive functions might provide the key for a general understanding of human problem solving. This might help to get deeper insights in the involved processes and to complement the classical Information Processing Theory.

6 Closing Remarks

In this chapter we have provided a selective overview of psychological theories of human problem solving, starting with the Gestaltist perspective, and following its adoption, transformation, and development within information processing theory. We hope to have shown the thread linking these apparently disconnected fields and provided a coherent perspective on human problem solving. In addition, we have highlighted two issues, ill-structured problem-solving and insight problem-solving, that test the limits of the information processing theory approach and suggest that additional, or even alternative, concepts and frameworks are necessary if we are to enhance our understanding of human thinking processes.

References

Anderson JR, Lebiere C (1998) The atomic components of thought. Erlbaum, Mahwah, NJ

Ash MG (1998) Gestalt psychology in German culture, 1890–1967: holism and the quest for objectivity Cambridge: Cambridge University Press, New York

Birch HG, Rabinowitz HS (1951) The negative effect of previous experience on productive thinking. J Exp Psychol 41:121–125

Bowden EM, Jung-Beeman M, Fleck J, Kounios J (2005) New approaches to demystifying insight. Trends Cogn Sci 9:322–328

Chronicle EP, MacGregor JN, Ormerod TC (2004) What Makes an Insight Problem? The Roles of Heuristics, Goal Conception, and Solution Recoding in Knowledge-Lean Problems. J Exp Psychol Learn Mem Cogn 30:14–27
Duncker K (1945) On problem-solving, vol 58. American Psychological Association INC, Washington
Ericsson KA, Simon HA (1993) Protocol analysis: verbal reports as data, revth edn. MIT Press, Cambridge, MA
Ernst GW, Newell A (1969) GPS: a case study in generality and problem solving. Academic Press, New York
Fodor JA, Pylyshyn Z (1988) Connectionism and cognitive architecture: a critical analysis. Cognition 28:3–71
Gigerenzer G, Hug K (1992) Domain-spcific reasoning: social contracts, cheating, and perspective change. Cognition 43:127–171
Gigerenzer G, Todd PM (2001a) Fast and frugal heuristics. In: Gigerenzer G, Todd PM (eds) Simple heuristics that make us smart. Oxford University Press, New York, pp 3–34
Gigerenzer G, Todd PM (2001b) Simple heuristics that make us smart. Oxford University Press, New York
Giunti M (1997) Computation, dynamics, and cognition. Oxford University Press, New York
Goel V (1995) Sketches of thought. MIT Press, Cambridge
Goel V (2005) Cognitive neuroscience of deductive reasoning. In: Holyoak KJ, Morrison RG (eds) The Cambridge handbook of thinking and reasoning. Cambridge University Press, Cambridge, pp 475–492
Goel V, Pirolli P (1992) The structure of design problem spaces. Cogn Sci 16:395–429
Grant ER, Spivey MJ (2003) Eye movements and problem solving: guiding attention guides thought. Psychol Sci 14:462–466
Holyoak KJ (1995) Problem solving. In: Smith EE, Osherson DN (eds) Thinking: an invitation to cognitive science, vol 3, 2nd edn. MIT Press, Cambridge, MA, pp 267–296
James W (1890/1950) The Principles of Psychology. vol 1. Dover, New York
Kaplan CA, Simon HA (1990) In search of insight. Cogn Psychol 22:374–419
Katona G (1940) Organizing and memorizing: studies in the psychology of learning and teaching. Columbia University, New York
Kershaw TC, Ohlsson S (2004) Multiple causes of difficulty in insight: the case of the nine-dot problem. J Exp Psychol Learn Mem Cogn 30(1):3–13
Knoblich G, Öllinger M (2006) The eureka moment. Sci Am Mind 10:38–43
Knoblich G, Ohlsson S, Haider H, Rhenius D (1999) Constraint relaxation and chunk decomposition in insight problem solving. J Exp Psychol Learn Mem Cogn 25(6):1534–1555
Koffka K (1935) Principles of gestalt psychology. Harcourt, Brace und World, New York
Köhler W (1921) Intelligenzprüfungen am Menschenaffen. Springer, Berlin
Köhler W (1925) The mentality of apes. Livewright, New York
Köhler W (1947) Gestalt psychology. Liveright, New York
Lovett MC, Anderson JR (1996) History of success and current context in problem solving: combined influences on operator selection. Cogn Psychol 31:168–217
Luchins AS (1942) Mechanization in problem solving – the effect of Einstellung. Psychol Monogr 54:1–95
Luchins AS, Luchins EH (1959) Rigidity of behavior: a variational approach to the effect of Einstellung. University of Oregon Books, Eugene, OR
Luchins AS, Luchins EH (1994) The water jar experiments and Einstellung effects: II Gestalt psychology and past experience. Gestalt Theory 16:205–259
MacGregor JN, Ormerod TC, Chronicle EP (2001) Information processing and insight: a process model of performance on the nine-dot and related problems. J Exp Psychol Learn Mem Cogn 27:176–201
Mai X-Q, Luo J, Wu J-H, Luo Y-J (2004) “Aha!” effects in a guessing riddle task: an event-related potential study. Hum Brain Mapp 22:261–271
Maier NRF (1930) Reasoning in humansI. On direction. J Comp Psychol 10:115–143

Maier NRF (1931) Reasoning in humans II. The solution of a problem and its appearance in consciousness. J Comp Psychol 12:181–194
Mayer RE (1992) Thinking, problem solving, cognition, 2nd edn. W. H. Freeman and Company, New York
Newell A (1990) Unified theories of cognition. Harvard University Press, Cambridge, MA, USA
Newell A, Simon HA (1972) Human problem solving. Prentice Hall, Englewood Cliffs, NJ
Newell A, Shaw JC, Simon HA (1958) Elements of a theory of human problem solving. Psychol Rev 65:151–166
Novick LR, Bassok M (2005) Problem Solving. In: Holyoak KJ, Morrison RG (eds) The Cambridge handbook of thinking and reasoning. Cambridge University Press, Cambridge, pp 321–350
Ohlsson S (1990) The mechanism of restructuring in geometry. In: Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, Lawrence Earlbaum Associates, Hillsdale, NJ, pp 237–244
Öllinger M, Knoblich G (2009) Psychological research on insight problem solving. In: Atmanspacher H, Primas H (eds) Recasting reality. Springer, Berlin, Heidelberg, pp 275–300
Öllinger M, Jones G, Knoblich G (2008) Investigating the effect of mental set on insight problem solving. Expe Psychol 55:269–282
Reitman WR (1964) Heuristic decision procedures, open constraints, and the structure of Ill-defined problems. In: Shelly MW, Bryan GL (eds) Human judgments and optimality. Wiley, New York, pp 282–315
Rittel HWJ, Webber MM (1974) Dilemmas in a general theory of planning. DMG-DRS J 8:31–39
Scheerer M (1963) Problem-Solving. Sci Am 208:118–128
Schooler JW, Engstler-Schooler TY (1990) Verbal overshadowing of visual memories: some things are better left unsaid. Cogn Psychol 22:36–71
Schooler JW, Ohlsson S, Brooks K (1993) Thoughts beyond words: when language overshadows insight. J Exp Psychol Gen 122:166–183
Simon HA (1973) The structure of ill structured problems. Artif Intell 4:181–201
Simon HA (1975) The functional equivalence of problem solving skills. Cogn Psychol 7:268–288
Thorndike EL (1911) Animal intelligence: experimental studies. Macmillian, New York
Wallas G (1926) The art of thought. Harcourt Brace Jovanovich, New York
Weisberg RW, Alba JW (1981) An examination of the alleged role of "fixation" in the solution of several "insight" problems. J Exp Psychol Gen 110:169–192
Wertheimer M (1912) Experimentelle Studien über das Sehen von Bewegungen. Zeitschrift für Psychologie 61:161–265
Wertheimer M (1925) Drei Abhandlungen zur Gestalttheorie. Verlag der Philosophischen Akademie, Erlangen
Wertheimer M (1959) Productive thinking. Harper, New York
Wickelgren WA (1974) How to solve it. Freeman, San Francisco

Heuristic Bias, Conflict, and Rationality in Decision-Making

Wim De Neys

Abstract Half a century of reasoning and decision-making research has shown that human thinking is often biased. People seem to over-rely on intuitions and gut feelings instead of on more demanding, deliberative reasoning when making decisions. The omnipresence of this bias has led to the questioning of human rationality. In this chapter I clarify that the crucial question for our view of human rationality is whether or not people detect that their intuitions conflict with more normative considerations when they are biased. In the first section I review recent conflict detection studies that started addressing this issue. The second section discusses the implications of the conflict detection work for the debate on human rationality. The key message is that focusing on the conflict detection process shows that people are far more rational and normative than their actual responses show.

1 Introduction

My dad runs a beer store. When buying a case of fancy Belgian beer, customers often ask whether they can buy a couple of matching glasses. My dad usually gets these glasses for free from his suppliers so he actually doesn't mind giving them away. However, he does not like to be easy on his customers and enjoys putting their decision-making skills to the test. When people ask him how much they owe him for the glasses, he tells them he is charging 5 euros for a glass but he also informs them that if they take a full box of six glasses instead of the one or two they asked for, they will get a 100% reduction. From a rational, economical point of view it is pretty obvious what people need to do. Two glasses will cost them 10 euros (2×5 euros $= 10$ euros). Six glasses would normally cost them 30 euros (6×5 euros $= 30$ euros) but thanks to the 100% reduction they will not be paying anything if they take the full box (100% of 30 euros is 30 euros, of course). This is a

W. De Neys
Lab Experimental Psychology, University of Leuven, Tiensestraat 102, 3000, Leuven, Belgium
e-mail: Wim.Deneys@psy.kuleuven.be

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_2,

very basic calculation that most elementary school children would have little trouble solving. Nevertheless, what my dad typically observes is that although he is catering to well-educated middle-class families, the vast majority of his customers decide to reject his offer. Even when he warns them that they are missing out on the 100% reduction, they still decide to stick to (and pay for!) the original number of glasses they asked for. Hence, people prefer to pay for glasses they could easily get for free. As my dad puts it, his customers' striking "failure to think" forces one to conclude that humans are ignorant, irrational beings.

Interestingly, the scientific study of human thinking seems to confirm my dad's observations. Since psychological studies of reasoning and decision-making started booming in the late 1950s, numerous studies have shown that in a wide range of reasoning and decision-making tasks, most educated adults are biased and fail to give the answer that is correct according to logic or probability theory (Evans and Over 1996; Kahneman et al. 1982). The general problem seems to be that reasoners over-rely on intuitions and gut feelings instead of on more demanding, deliberative reasoning when making decisions (Evans 2003; Kahneman 2002). Although this intuitive or so-called "heuristic" thinking might sometimes be useful, it will often cue responses that are not warranted from a normative point of view. Consequently, people's reasoning and decision-making is often biased.

It is not hard to see how such intuitive or heuristic thinking is biasing my dad's customers in his store. Intuitively, people's gut feeling might simply be telling them that by offering an additional reduction my dad is trying to persuade them to buy more than they asked for. In general, such a heuristic might be a useful tool to prevent falling prey to sales tricks. However, in my dad's store this mere intuitive reasoning is costing people good money. Hence, the point is not that heuristics or intuitions are necessarily bad. The point is rather that during reasoning and decision-making it is crucial to check whether one's intuitions conflict with more normative considerations. As my dad would claim, the omnipresence of heuristic bias suggests that people are not very good at detecting such conflicts.

The conflict detection process is a key component of any theory of reasoning and decision-making. Unfortunately, the process is poorly understood and there are some quite different views on its efficiency. Consistent with my dad's view, for example, a number of authors have argued that conflict detection during thinking is quite unsuccessful (e.g., Evans 1984; Kahneman and Frederick 2002). According to these authors, the widespread heuristic bias can be attributed to a failure to monitor our intuition. Because of lax monitoring people would simply fail to detect that the intuitive response conflicts with more normative considerations. Bluntly put, people would be biased because they do not notice that their intuition is wrong.

However, others have suggested that conflict detection during thinking is actually pretty flawless (e.g., Epstein 1994; Sloman 1996). According to these authors, there is nothing wrong with the detection process. People do notice that the intuitive response conflicts with more normative considerations. The problem, however, is that despite this knowledge they will not always manage to inhibit and discard the tempting intuitive beliefs. Thus, people "behave against their better judgment" (Denes-Raj and Epstein 1994, p. 1) when they give an unwarranted heuristic

response: they detect that they are biased but simply fail to block the biased response. In sum, in this view biased decisions are attributed to an inhibition failure rather than a conflict detection failure per se (see also Houdé 2007).

Clarifying the efficiency of the conflict detection process and the resulting nature of the heuristic bias is paramount for the development of reasoning and decision-making theories. The issue also has far-stretching implications for our view of human rationality. If the popular bias-as-detection-failure view is right and reasoners do not detect that their heuristic response is wrong, this implies that reasoning errors are indeed quite "dumb." The second view, however, implies that people's errors are less ignorant. If people detect that their intuitive response is not fully warranted, this implies that people did not simply neglect the normative considerations. Contrary to my dad's conclusion, this would suggest that people are no mere heuristic thinkers and might be more rational than their actual responses show.

The problem, however, is that it is hard to decide between the alternative views based on traditional reasoning data (Evans 2007, 2008a,b). Recently, however, there have been some initial attempts to break the stalemate. A number of studies started developing processing measures of conflict detection during reasoning. In the following section I will briefly review this work. In a final section I will discuss the implications of the findings for the debate on human rationality in more detail.

2 Conflict Detection Studies

2.1 To Detect or Not to Detect?

De Neys and Glumicic (2008) recently presented one of the first studies that explicitly focused on an empirical test of the efficiency of the conflict detection process during thinking. They pointed out that the classic claims about the detection process were typically anecdotal in nature. (Epstein 1994; Denes-Raj and Epstein 1994; Epstein and Pacini 1999), for example, repeatedly noted that when picking an erroneous answer his participants spontaneously commented that they did "*know*" that the response was wrong but stated they picked it because it "*felt*" right. Such comments do seem to suggest that people detect that their intuition conflicts with normative considerations. The problem, however, is that spontaneous selfreports and anecdotes are no hard empirical data. This is perhaps best illustrated by the fact that Kahneman (2002, p. 483) also refers to "casual observation" of his participants to suggest that only in "some fraction of cases, a need to correct the intuitive judgements and preferences will be acknowledged." Therefore, in a first experiment De Neys and Glumicic decided to adopt a thinking aloud procedure (e.g., Ericsson and Simon 1993). The thinking aloud procedure has been designed to gain reliable information about the course of cognitive processes. Participants are simply instructed to continually speak aloud the thoughts that are in their head as

they are solving a task. Thinking aloud protocols have been shown to have a superior validity compared to interpretations that are based on retrospective questioning or people's spontaneous remarks (Payne 1994).

De Neys and Glumicic (2008) asked their participants to solve problems that were modeled after Kahneman and Tversky's classic (Kahneman and Tversky 1973) base-rate neglect problems. These base-rate neglect problems are among the most (in)famous tasks in the field. In the problems people first get information about the composition of a sample (e.g., a sample with 995 females and 5 males). People are also told that short personality descriptions are made of all the participants and they will get to see one description that was drawn randomly from the sample. Consider the following example:

A psychologist wrote thumbnail descriptions of a sample of 1,000 participants consisting of 995 females and 5 males. The description below was chosen at random from the 100 available descriptions.

Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer.

Which one of the following two statements is most likely?

(a) Jo is a man
(b) Jo is a woman

From a normative point of view, given the size of the two groups in the sample, it is more likely that a randomly drawn individual will be a female. However, intuitively many people will be tempted to respond that the individual is a male based on stereotypical beliefs cued by the description ("Jo is an engineer and drinks beer").

The crucial question for De Neys and Glumicic was whether verbal protocols would indicate that when people selected the intuitive response option ("a. Jo is a man") they at least referred to the group size information during the reasoning process (e.g., " … because Jo's drinking beer and loud I guess Jo'll be a guy, *although there were more women* …"). In this task such basic sample size reference during the reasoning process can be considered as a minimal indication of successful conflict monitoring. It indicates that this information is not simply neglected.

Results were pretty straightforward. People who gave the correct response typically also referred to the base-rate information and reported they were experiencing a conflict (e.g., "… it sounds like he's a guy, *but because they were more women*, Jo must be female so I'll pick option b …"). However, people who gave the intuitive response hardly ever (less than 6 % of the cases) mentioned the base-rate information (e.g., a typical protocol would read something like " … This person is a guy … drinks, listens to loud music … yeah, must be a guy … so I'll pick a … "). Hence, consistent with my dad's claims and the error-as-detection-failure view, the verbal protocols seemed to indicate that people are indeed mere intuitive reasoners who do not detect that they are biased.

De Neys and Glumicic noted, however, that it could not be excluded that conflict detection was successful at a more implicit level. It might be that the conflict detection experience is not easily verbalized. People might notice that there is something

wrong with their intuitive response but they might not always manage to put their finger on it. Such more implicit conflict detection would still indicate that people detect that their response is not fully warranted, of course. To capture such implicit detection De Neys and Glumicic also presented participants with a surprise recall test. After a short break following the thinking-aloud phase participants were asked to answer questions about the group sizes in the previous reasoning task. Participants were not told that recall would be tested while they were reasoning but De Neys and Glumicic reasoned that the detection of the conflict should result in some additional scrutinizing of the normative base-rate information. This deeper processing of the base-rate information should subsequently benefit recall.

To validate the recall hypothesis participants were also presented with additional control problems. In the classic base-rate problems the description of the person is composed of common stereotypes of the smaller group so that the normative response cued by the base-rates and the intuitive response that is cued by the description disagree. In addition to these classic problems De Neys and Glumicic also presented problems in which the base-rates and description both cued the same response. In these *congruent* problems the description of the person was composed of stereotypes of the *larger* group. Hence, contrary to the classic (i.e., *incongruent*) problems the intuitive response did not conflict with more normative considerations and the response could be rightly based on mere intuitive processing. For a reasoner who neglects the base-rates and does not detect the conflict on the classic problems, both types of problems will be completely similar and base-rate recall should not differ. However, if one does detect the conflict, the deeper processing of the base-rates in case of a conflict should result in better recall for the classic problems than for the congruent control problems.

Recall results showed that participants had indeed little trouble recalling the base-rates of the classic conflict problems. People easily remembered which one of the two groups in each problem was the largest. On the congruent control problems, however, recall performance was merely at chance level. Interestingly, the superior recall was obvious even for those people who never mentioned the base-rates while thinking-aloud and failed to solve any of the presented classic conflict problems correctly. Since the only difference between the classic and control problems was the conflicting nature of the base-rates and description, De Neys and Glumicic concluded that people had little difficulty in detecting the conflict per se.

In an additional experiment De Neys and Glumicic examined the conflict detection issue further by introducing a "moving window" procedure (e.g., Just et al. 1982). In the experiment the base-rates and the description were presented separately. First, participants saw the base-rate information on a computer screen. Next, the description and question were presented and the base-rates disappeared. Participants had the option of visualizing the base-rates afterwards by holding a specific button down. Such base-rate reviewing can be used as an additional conflict detection index. De Neys and Glumicic explained their recall findings by assuming that when people detect that the description conflicts with the previously presented base-rates they will spend extra time scrutinizing or "double checking" the base-rates. With the "moving window" procedure the time spent visualizing the

base-rates can be used as a measure of this reviewing tendency. If conflict detection is indeed successful, people should show a stronger tendency to visualize the base-rates when solving classic incongruent vs. congruent control problems. This is exactly what De Neys and Glumicic observed. Once again the stronger base-rate reviewing was present for the least-gifted reasoners in the sample who consistently gave the intuitive response on all presented incongruent problems.

2.2 To the Brain and Beyond

In a further attempt to clarify the nature of heuristic bias (De Neys et al. (2008)) decided to focus on the neural basis of conflict detection and response inhibition during thinking. They noted that numerous imaging studies established that conflict detection and actual response inhibition are mediated by two distinct regions in the brain. Influential work in the cognitive control field (e.g., Botvinick et al. 2004; Ridderinkhof et al. 2004), for example, showed that detection of an elementary conflict between competing responses is among the functions of the medial part of the frontal lobes, more specifically the Anterior Cingulate Cortex (ACC). While the ACC signals the detection, correct responding and actually overriding the erroneous, prepotent response has been shown to depend on the recruitment of the more lateral part of the frontal lobes (more specifically the right lateral prefrontal cortex or RLPFC).

De Neys, Vartanian, and Goel therefore suggested that turning to the brain might help to address the dispute about the nature of heuristic bias. Solving classic decision-making problems that cue a salient but inappropriate intuitive response requires that reasoners detect that the intuitive response conflicts with normative considerations, first. In addition, the intuitive responses will need to be successfully inhibited. If the ACC and RLPFC mediate this conflict detection and inhibition process, respectively, correct reasoning should be associated with increased activation in both areas. De Neys, Vartanian, and Goel reasoned that the crucial nature of the intuitive bias could be clarified by contrasting ACC and RLPFC activation for intuitive and normative responses. The bias-as-inhibition-failure and bias-as-detection-failure views make differential predictions with respect to the activation of the conflict detection region. If De Neys and Glumicic's initial behavioral findings were right and people at least detect that the intuitive response conflicts with more normative considerations, the ACC should be activated whether or not people are biased. However, if biased decisions arise because people fail to detect that the intuitive response is inappropriate, people will not detect a conflict when they give an intuitive response and consequently the ACC should not be activated.

De Neys, Vartanian, and Goel tested these predictions in an fMRI study in which participants were asked to solve base-rate problems while the activation of the ACC and RLPFC was monitored. As expected, results showed that for trials in which people selected the correct base-rate response on the classic, incongruent problem versions, both the conflict detection (ACC) and inhibition region (RLPFC) showed

increased activation. When people were biased and selected the intuitive response on these problems, the RLPFC inhibition region was not recruited. The conflict detection ACC region, however, did show clear activation when the intuitive response was selected. On congruent control trials in which the cued intuitive and normative response did not conflict, the ACC was not activated.

In sum, De Neys, Vartanian, and Goel's crucial finding was that biased and unbiased responses on the classic base-rate problems only differed in RLPFC recruitment. Solving incongruent problems did engage the ACC region but the activation did not differ for intuitive or base-rate responses. Consistent with De Neys and Glumicic's behavioral findings this suggested that the intuitive bias should not be attributed to a detection failure but rather to an inhibition failure.

2.3 The Effortless Nature of Conflict Detection

Taken together the De Neys and Glumicic (2008) and De Neys, Vartanian, and Goel studies (2008) supported the view of authors such as Epstein (1994) who claimed that conflict detection during thinking is pretty flawless. However, the absence of any verbally expressed conflict experience suggested that the popular characterization of this process as an explicitly experienced struggle in which people are actively deliberating between two different options ("I know it's wrong but it feels right") is not very accurate. Hence, Franssens and De Neys (2009) recently argued that the conflict detection process itself might be better conceived as an intuitive process that simply warns people that more deliberate reasoning is required (see also Evans, in press). Although the conflict detection would suffice to inform people that their heuristic conclusion is not fully warranted and needs to be scrutinized, it would not guarantee that further deliberate reasoning is actually engaged in to override and inhibit the heuristic response. Bluntly put, it looks like people intuitively feel that "something" is wrong but, without more demanding deliberate thinking, cannot exactly specify what.

Franssens and De Neys (2009) presented a straightforward experiment to test the claim that conflict detection is an intuitive process. One of the key characteristics of intuitive, implicit processing is that it is effortless and does not draw on people's limited executive working memory resources that are required for controlled processing (e.g., Moors and De Houwer 2006). Franssens and De Neys therefore decided to burden these executive resources during reasoning. In their study participants were asked to memorize spatial dot patterns while they were trying to solve base-rate problems. This dot memorization task had been previously shown to burden the executive resources (Miyake et al. 2001). Franssens and De Neys reasoned that if conflict detection during thinking was indeed intuitive, it should not be affected by the executive memorization load. The efficiency of the conflict detection process was measured by presenting the participants with the surprise base-rate recall task that was introduced in the De Neys and Glumicic (2008) studies. Results showed that reasoning performance per se decreased under memorization load.

Participants gave more heuristic responses when their executive resources were burdened. However, the recall performance was not affected. Even under load base-rate recall was still better for classic incongruent than for congruent control problems and the percentage correct recall for the incongruent problems did not differ under load and no-load conditions. Hence, the study nicely supported the characterization of conflict detection as a flawless and intuitive process.

3 Implications for the Rationality Debate

The studies reviewed above suggest that people are quite good at detecting the conflict between cued heuristic intuitions and more normative considerations when solving classic decision-making problems. Although people's responses are typically biased they do seem to have an intuitive gut feeling that is telling them that their heuristic answer is not fully warranted. Even though it is hard for people to verbalize this intuitive conflict feeling, its flawless manifestation indicates that normative considerations are not simply neglected. If people were not to know the normative principles (e.g., the fact that base-rates matter) or would not consider these normative principles to be relevant, there would simply be no conflict to be detected in the first place and congruent and incongruent problem versions should be processed in the exact same manner. Clearly, conflict can only occur when both the intuitive response and normative considerations are taken into account during thinking. The fact that people are particularly sensitive to the presence of this conflict when solving classic decision-making problems implies that people are no mere heuristic thinkers who simply neglect normative considerations. In this section I will try to clarify that this point has some profound implications for the debate about the rationality of the human species (e.g., Stanovich and West 2000; Stein 1996).

The so-called "rationality debate" has raged through the reasoning and decision-making field for more than four decades without clear solution. In essence, the debate centers around two related questions: (a) whether human reasoning is rational and (b) whether the traditional normative systems (such as logic and probability theory) against which the rationality of our inferences and decisions are measured are actually valid. The initial findings in the 1960's that pointed to the omnipresence of heuristic bias led some theorists to question the rationality of the human species (e.g., Wason 1968, 1983; see Evans 2002, for a nice review). Just like my dad in his store, these theorists concluded that people's widespread failure to reason in line with the logical or probabilistic norm indicated that humans are irrational beings. However, later on this pessimistic conclusion was rejected by theorists who started questioning the validity of the classic norms. Bluntly put, it was argued that if the vast majority of well-educated, young adults fail to solve a simple reasoning task, this might indicate that there is something wrong with the task scoring norm rather than with the participants. The basic point of these authors was that people might interpret the tasks differently and adhere to other norms than the classic ones (e.g., Hertwig and Gigerenzer 1999; Oaksford and Chater 1998; Todd and

Gigerenzer 2000). For example, in the base-rate problems participants might interpret the task as a simple social classification task and would therefore not keep track of the base-rate information. These authors clarified that the rationality of our behavior depends on the goals we try to fulfill. If our goal is making a social classification judgment, neglecting the base-rates is the rational thing to do and cannot be considered a bias. Hence, according to this "alternative norm" view, people's behavior in the classic reasoning and decision-making experiments is perfectly rational but has simply been measured against the wrong standards.

One might note that the opposite rationality views are trading-off rationality and norm validity. People like my dad take the validity of the classic norms for granted and conclude that the failure to reason in line with these norms points to human irrationality. The "alternative norm" view on the other hand saves human rationality but at the cost of the validity of the classic norms. I believe that studying the conflict detection process during thinking presents an opportunity to resolve this debate and unify the two views. The initial conflict detection data that I reviewed suggest that both human rationality and the validity of the classic norms can be saved. If people were really to interpret classic reasoning and decision-making tasks as social classification tasks and were to believe that normative considerations such as sample sizes do not matter, their task processing should not be affected by the presence of a conflict between cued social intuitions and the very same normative principles. Hence, contrary to the "alternative norm" view this indicates that people do not consider the classic norms to be irrelevant. On the other hand, the fact that people pick up this conflict shows that they take normative considerations into account and are no mere intuitive thinkers. In sum, people might not always manage to reason in line with the classic norms but this does not imply that they do not know the norms or consider them to be irrelevant. The initial conflict detection studies suggest that all reasoners are at least trying to adhere to the classic norms and detect that their intuition is not warranted.

4 Caveats and Conclusion

In this chapter I wanted to highlight a new research framework in the reasoning and decision-making field that started focusing on the efficiency of the conflict detection process during thinking. Needless to say, this framework is still in its infancy and the initial findings and conclusions need to be interpreted with some caution. Clearly, the work will need to be validated and generalized in future studies. However, I hope to have clarified the potential and importance of this line of research. The key point is that a failure to characterize the conflict detection process during thinking is bound to bias any conclusions about human rationality or the validity of the classic norms. The initial conflict processing data indicates that people are pretty good at detecting their bias. Contrary to popular views in the decision-making field and the opinion of at least one Belgian beer expert, this suggests that people are far more rational and normative than their biased answers suggest.

Acknowledgments My work is funded by the Fund for Scientific Research Flanders (Fonds Wetenschappelijk Onderzoek Vlaanderen). I would like to thank my dad, Hubert De Neys, for passing on his interest in human reasoning and decision-making. I promised him that I would let the interested readers know that they can pay a virtual visit to his store at http://www.deneys-asselman.be/ .

References

Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci 12:539–546

Denes-Raj V, Epstein S (1994) Conflict between intuitive and rational processing: When people behave against their better judgement. J Pers Soc Psychol 66:819–829

De Neys W, Glumicic T (2008) Conflict monitoring in dual process theories of reasoning. Cognition 106:1248–1299

De Neys W, Vartanian O, Goel V (2008) Smarter than we think: when our brains detect that we are biased. Psychol Sci 19:483–489

Epstein S (1994) Integration of the cognitive and psychodynamic unconscious. Am Psychol 49:709–724

Epstein S, Pacini R (1999) Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. In: Chaiken S, Trope Y (eds) Dual process theories in social psychology. Guilford Press, New York, pp 462–482

Ericsson KA, Simon HA (1993) Protocol analysis: verbal reports as data. MIT Press, Cambridge, MA

Evans J St B T (1984) Heuristic and analytic processing in reasoning. Br J Psychol 75:451–468

Evans J St B T (2002) Logic and human reasoning: an assessment of the deduction paradigm. Psychol Bull 128:978–996

Evans J St B T (2003) In two minds: dual process accounts of reasoning. Trends Cogn Sci 7:454–459

Evans JStBT (2007) On the resolution of conflict in dual process theories of reasoning. Think Reas 13:321–329

Evans JStBT (2008a) Dual-processing accounts of reasoning, judgement and social cognition. Annu Rev Psychol 59:255–278

Evans JStBT (2008b) How many dual process theories do we need: one, two or many? In: Evans JStBT, Frankish K (eds) In two minds: dual processes and beyond. Oxford University Press, Oxford

Evans JStBT, Over DE (1996) Rationality and reasoning. Psychology Press, Hove, UK

Franssens S, De Neys W (2009) The effortless nature of conflict detection during thinking. Think Reas 15:105–128

Hertwig R, Gigerenzer G (1999) The "conjunction fallacy" revisited: how intelligent inferences look like reasoning errors. J Behav Decis-making 12:275–305

Houdé O (2007) First insights on "neuropedagogy of reasoning". Think Reas 13:81–89

Just MA, Carpenter PA, Wooley JD (1982) Paradigms and processes in reading comprehension. J Exp Psychol Gen 111:228–238

Kahneman D (2002) Maps of bounded rationality: a perspective on intuitive judgement and choice. Nobel Prize Lecture. Retrieved January 11, 2006, from http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf

Kahneman D, Frederick S (2002) Representativeness revisited: attribute substitution in intuitive judgement. In: Gilovich T, Griffin D, Kahneman D (eds) Heuristics and biases: the psychology of intuitive judgement. University press, Cambridge, pp 49–81

Kahneman D, Slovic P, Tversky A (1982) Judgement under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, MA

Kahneman D, Tversky A (1973) On the psychology of prediction. Psychol Rev 80:237–251
Miyake A, Friedman NP, Rettinger DA, Shah P, Hegarty M (2001) How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. J Exp Psychol Gen 130:621–640
Moors A, De houwer J (2006) Automaticity: a theoretical and conceptual analysis. Psychol Bull 132:297–326
Oaksford M, Chater N (1998) Rationality in an uncertain world: essays on the cognitive science of human reasoning. Psychology Press, Hove, UK
Payne JW (1994) Thinking aloud: Insights into information processing. Psychol Sci 5:241–248
Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal cortec in cognitive control. Science 306:443–447
Sloman SA (1996) The empirical case for two systems of reasoning. Psychol Bull 119:3–22
Stanovich KE, West RF (2000) Individual differences in reasoning: Implications for the rationality debate. Behav Brain Sci 23:645–726
Stein E (1996) Without good reason: The rationality debate in philosophy and cognitive science. Oxford University Press, Oxford, England
Todd PM, Gigerenzer G (2000) Precis of simple heuristics that make us smart. Behav Brain Sci 23:727–780
Wason PC (1968) Reasoning about a rule. Q J Exp Psychol 60:273–281
Wason PC (1983) Realism and rationality in the selection task. In: Evans JStBT (ed) Thinking and reasoning. Routledge, London, pp 45–75

Analogical Processes in Human Thinking and Learning

Dedre Gentner and Julie Colhoun

Abstract Much of humankind's remarkable mental aptitude can be attributed to analogical ability – the ability to perceive and use relational similarity. In this chapter, we present an overview of analogy and describe its component processes, including structural alignment and inference projection, evaluation, schema abstraction and re-representation. We discuss how these component processes lead to learning and the generation of new knowledge, and review evidence that suggests that greater use of analogy during learning can improve relational retrieval and transfer.

1 Introduction

Similarity and association are two great forces of mental organization that hold across species. Although humans probably experience the same kinds of intuitive connections as do hamsters, our species also experiences a more sophisticated form of each of these two forces: namely, analogy (a selective form of similarity) and causation (a selective form of association). In this chapter we focus on analogy – the perception of like relational patterns across different contexts. The ability to perceive and use purely relational similarity is a major contributor – arguably *the* major contributor – to our species' remarkable mental agility (Gentner 2003; Gentner and Christie 2008; Kurtz et al. 1999; Penn et al. 2008). Understanding how it works is thus important in any account of "why we're so smart" (Gentner 2003).

A good analogy both reveals common structure between two situations and suggests further inferences. For example, discussions of cell biology sometimes explain cell metabolism by analogy with a fire:

D. Gentner (✉) and J. Colhoun
Department of Psychology, Northwestern University, Evanston, 60208, IL, USA
e-mail: gentner@northwestern.edu

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_3,

- A fire consumes fuel using oxygen, thereby releasing energy; it releases carbon dioxide and water.
- Likewise, a cell's mitochondria obtain energy from glucose using oxygen, in a process called oxidation.

This analogy highlights the common relational structure: that cell metabolism can be seen as the burning of fuel, and fire as a form of oxidation. It also invites the (correct) inference that cell metabolism releases water and carbon dioxide. In such explanatory analogies, a familiar situation, referred to as the *base* or *source* analog, is used as a model by which to understand and draw new inferences to the unfamiliar situation or *target*. Recent research has also focused on another use of analogy in learning – namely, to reveal the common structure between two situations, neither of which needs to have been fully understood before the comparison. In this paper, we begin by presenting an overview of analogy and its component processes. We then discuss each component process in greater detail.

2 Analogical Processes

Theories of analogy distinguish the following processes: (1) *retrieval*: given some current situation in working memory, a prior similar or analogous example may be retrieved from long-term memory; (2) *mapping*: given two cases in working memory, mapping consists of *aligning* their representational structures to derive the commonalities and *projecting inferences* from one analog to the other. Mapping is followed by (3) *evaluation* of the analogy and its inferences and often by (4) *abstraction* of the structure common to both analogs. A further process that may occur in the course of mapping is (5) *re-representation*: adaptation of one or both representations to improve the match. We begin with the processes of mapping through re-representation, reserving retrieval for later.

2.1 Mapping

Mapping is the heart of analogy, and, not surprisingly, it has been a central focus in analogy research. According to Gentner's (Gentner 1983, 1989; Gentner and Markman 1997) structure-mapping theory, analogical mapping is the process of establishing a *structural alignment* between two represented situations and then projecting *inferences*. The theory assumes structured representations in which the elements are connected by labeled relations, and higher-order relations (such as causal relations) connect first-order statements (see Falkenhainer et al. 1989; Markman 1999). During the alignment process (as amplified below), possible matches are first found between individual elements of the two represented situations; then these matches are combined into structurally consistent clusters, and

finally into an overall mapping. The resulting alignment consists of an explicit set of correspondences between the sets of representational elements of the two situations, with an emphasis on matching relational predicates. As a natural outcome of the alignment process, candidate inferences are projected from the base to the target. These inferences are propositions connected to the common system in one analog, but not yet present in the other. An example from our earlier analogy is the inference that cell metabolism produces CO_2 and water as by-products.

The alignment process is guided by a set of tacit constraints that lead to structural consistency: (a) there must be *one-to-one correspondence* between the mapped elements in the target and base, and (b) there must be *parallel connectivity,* such that the arguments of corresponding predicates also correspond. A further assumption is the *systematicity principle:* in selecting among possible interpretations of an analogy, a system of relations that are connected by higher-order constraining relations (such as causal relations) is preferred over an equal number of independent matches. This principle guides the selection of an alignment, such that the more systematic of two possible alignments will be chosen. The systematicity principle reflects an implicit preference for coherence and predictive power in analogical processing. Thus, a base domain that possesses a richly linked system of relations will yield candidate inferences by completing the corresponding structure in the target (Bowdle and Gentner 1997).

The mapping process has been operationalized in the Structure Mapping Engine (SME; Falkenhainer et al. 1989), a computational model that instantiates Gentner's (1983) structure-mapping theory. This system operates in a local to global fashion, first finding all possible local matches between the elements of two potential analogs. It combines these into structurally consistent clusters, and then combines the clusters (called kernels) into the largest and most deeply connected system of matches. As noted above, other propositions connected to the common system in one analog become candidate inferences about the other analog. Finally, SME generates a structural evaluation of the match (see Forbus et al. 1995, for details).

The claim that analogical processing is symmetric at the outset might seem surprising, given the strong directionality of many analogies. For example, the statement "My surgeon is like a butcher" conveys a very different set of inferences from "My butcher is like a surgeon." This strong directionality has led some researchers to suggest that the processing of metaphors (Glucksberg et al. 1997) and analogies (Greiner 1988; Hummel and Holyoak 1997) is asymmetric from the start. However, according to structure-mapping, although inference projection is directional, it is guided by an initial alignment that is symmetric.

To test whether the initial stage is indeed symmetric, Wolff and Gentner (2000) and Gentner and Wolff (1997) investigated the processing of highly directional metaphors. These metaphors, like many of the metaphors used in psychological research, were essentially analogies, in that they conveyed a matching relational system: e.g., "Some jobs are jails." Furthermore, they were highly directional (Ortony 1979): "Some jobs are jails" is not at all the same as saying (quite incomprehensibly) "Some jails are jobs." In one series of studies, Wolff and Gentner (in preparation) gave participants these forward and reversed metaphors in a

speeded task (in either forward or reversed direction) and asked them to press either "comprehensible" or "not comprehensible." The results suggested that metaphor processing is symmetrical in the initial stages. At 600 ms, participants found forward and reversed metaphors equally comprehensible; not until roughly 1,200 ms did they show higher comprehension of the forward than of the reversed metaphors. This result did not stem from inability to process meaning at 600 ms, because even at this early deadline, participants rejected scrambled metaphors ("Some butchers are flutes") as incomprehensible and accepted literally true statements ("Some birds are robins") as comprehensible. This pattern of early symmetry followed by later directionality is in accord with the structure-mapping prediction of an initial symmetric alignment followed by later directional inferences from base to target (Gentner 1983, 1989; Falkenhainer et al. 1989).

2.2 *Structural Alignment in Similarity and Analogy*

The framework originally developed for analogy extends to literal similarity, as demonstrated by a series of studies at the University of Illinois in the 1990s (Gentner and Markman 1995, 1997; Goldstone et al. 1991; Markman and Gentner 1993a,b,c; Medin et al. 1993). The distinction between analogy and literal similarity can be thought of within a similarity space defined by the degree of object–attribute similarity and the degree of relational similarity, as shown in Fig. 1 (Gentner & Markman 1997). Analogy and literal similarity lie on a continuum based on the degree of object–attribute similarity between the items being compared. When a comparison exhibits a high degree of relational similarity with very little attribute similarity, we consider it an analogy. As the amount of attribute similarity increases, the comparison becomes one of literal similarity. This is not merely a matter of terminology. Literal similarity matches are easier to make (and

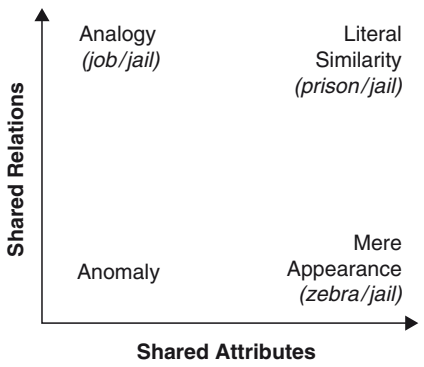

Fig. 1 Similarity space defined by the degree of object–attribute similarity and the degree of relational similarity. Adapted from Gentner and Markman (1997)

more accessible to novices and children) than analogies because the alignment of relational structure is supported by object matches.

Recent developmental research has shown that young learners can take advantage of close literal similarity matches to gain the beginnings of relational insight. Even a highly concrete literal similarity match involves an alignment of the relational structure, and that carrying out an "easy" literal match can render learners to better carry out a difficult relational match. For example, Loewenstein and Gentner (2001) give children (aged 3½) a challenging search task (DeLoache 1987). Children watched the experimenter hide a toy in a small model room (the Hiding room), and then tried find the toy hidden "in the same place" in a second model room (the Finding room). The two rooms contained the same type of furniture (bed, table, etc.) in the same configuration, but were rather dissimilar in the specific shapes of their furniture, making the mapping task difficult for these young children. Before engaging in the task, all the children were shown the Hiding room along with another very similar room (identical except for color). Half the children saw the two rooms together and were encouraged to compare them; the other half talked about each room separately. Children in the comparison condition were significantly more likely to correctly locate the toy in the Finding room than those who saw the rooms separately.

These findings have two important implications. First, the finding that even comparing close literally similar examples can promote highlighting of the common relational structure is further evidence that "similarity is like analogy" in promoting a structural alignment (Gentner and Markman 1995). Second, the finding that an easily aligned literal match can bootstrap young children to a more distant relational mapping offers a route by which children's ordinary experiential learning can gradually lead them to the discovery of analogical matches (Gentner and Medina 1998).

This progressive alignment process can help to dispel the mystery of how abstract ideas can arise from experience. Consider the example of monotonic change as it might first be learned by a child in a highly concrete context, such as the descending heights of a "Daddy Mommy Baby" set of dolls. The relational structure of descending size is at first implicit and embedded in the specific family context. At this stage the child would not recognize that the same structure occurs in, say, a set of bowls of decreasing diameter. But if the child is given a close match – say, a different set of descending-size dolls – then the obvious similarities will prompt an alignment process and help to guide it. Miraculously, even such a close alignment can elevate the salience of the common relational structure, thereby potentiating a subsequent more distant match, such as that between the dolls and the bowls. If this process continues – with each new analog clarifying and refining the common structure further – the result can become steadily more abstract (see Kotovsky and Gentner 1996, as discussed later, for an example). These close alignments, so mundane as to be nearly invisible to adults, can nonetheless accumulate, resulting in significant gains in learning.

Literal similarity supports the mapping process, but in some cases, object matches among elements of compared items can be a pitfall. Specifically, when items are

cross-mapped (Gentner and Toupin 1986) – that is, when similar (or identical) objects play different roles in the relational structure of each analog – the object match can be difficult to ignore. For example, if one analog describes a dog chasing a cat and the other describes a cat chasing a mouse, the cat is said to be cross-mapped. Such cross-mappings can be compelling for children and novices, especially if the object matches are rich and distinctive (Gentner and Rattermann 1991; Paik and Mix 2006). In general, the deeper and better-established the relational structure (as comes with expertise), the better a cross-mapping can be withstood (Gentner and Rattermann 1991; Gentner and Toupin 1986; Markman and Gentner 1993c).

2.3 Systematicity

The role of relational structure in analogical processing is more specific than a simple preference for relational commonalities over attribute or object matches. Ultimately, what makes comparison so revealing is that (for whatever reason) people like to find connected relational structure. Thus, the analogical interpretation process seeks matches that consist of interconnected systems of relations. As noted above, this preference for systematic interpretations is known as the systematicity principle. The claim that comparison promotes systems of interrelated knowledge is crucial to analogy's viability as a reasoning process. If the comparison process were to generate only isolated feature matches, there would be no natural basis for constraining which inferences are derived from the match.

In order to test whether systematicity constrains analogical matching, Clement and Gentner (1991) showed participants analogous scenarios and asked them to judge which of two lower-order assertions shared by the base and target was most important to the match. Participants chose the assertion that was connected to matching causal antecedents – their choice was based not only on the goodness of the local match, but also on whether it was connected to the larger matching system. Thus, matching lower-order relations such as (causal antecedents) that are interconnected by higher-order relations yield a better analogical match than an equal number of matching relations that are unconnected to each other.

A parallel result was found for inference projection: people were more likely to import a fact from the base to the target when it was connected to the common system (Clement and Gentner 1991; Markman 1997). In analogical matching, people are not interested in isolated coincidental matches; rather, they seek causal and logical connections, which give analogy its inferential power. The critical finding that systematicity guides inference also carries over to similarity comparisons. Bowdle and Gentner (1997) gave participants pairs of similar scenarios (without distinguishing base and target) and asked for inferences. Participants preferred to make inferences from a systematic structure to a less systematic structure and also judged comparisons to be more informative in this direction than the reverse. Similarly, Heit and Rubinstein (1994) demonstrated that people make stronger inferences when the kind of property to be inferred (anatomical or behavioral) matches the kind of similarity between the animals

(anatomical or behavioral). For instance, people make stronger behavioral inferences from tuna to whales (because both share behavioral capacities related to swimming) than from bears to whales, but stronger anatomical inferences from whales to bears (because both are mammals and therefore share an internal system of anatomical relations). These findings are consistent with the claim that people are strongly influenced by systematicity when drawing inferences from comparisons.

2.4 Evaluation

Although we have already alluded to evaluation in the course of this discussion, a few further points require mention. Specifically, evaluating an analogy and its inferences involves several kinds of judgment. One criterion is structural soundness: whether the alignment and the projected inferences are structurally consistent. With respect to particular candidate inferences, this translates to the amount of structural support the alignment provides for the inference. In addition to structural support, Forbus et al. (1997) suggest that another criterion may be the amount of new knowledge generated. That is, inferences that potentially yield a significant gain in new knowledge may be desirable (even if somewhat risky), especially when brainstorming or dealing with unfamiliar domains.

Another criterion, of course, is the factual validity of the projected inferences in the target. Because analogy is not a deductive mechanism, these candidate inferences are only hypotheses; their factual validity is not guaranteed by their structural consistency and must be checked separately. Thus, this type of evaluation may involve other reasoning processes such as causal reasoning from existing knowledge in the target. A fourth criterion, which applies in problem-solving situations, is pragmatic relevance – whether the analogical inferences are relevant to the current goals (Holyoak and Thagard 1989). An analogy may be structurally sound and yield true inferences, but still fail the relevance criterion if it does not bear on the problem at hand. A related criterion, discussed by Keane (1996), is the adaptability of the inferences to the target problem.

The evaluation of inferences and of the whole analogy can mutually influence one another. Evaluation of particular inferences contributes to the larger evaluation of the analogy, and if particular inferences are clearly false, the analogy loses force. Likewise, if the analogy consists of a poor structural match, the inferences garner less confidence.

3 Learning

There are three main ways in which an analogy can lead to learning and representational change in one or both analogs: projection of candidate inferences, schema abstraction – in which the highlighted relational structure is extracted and

stored – and re-representation of the constituent predicates of the analogs (Clement and Gentner 1991; Clement 1988; Holyoak and Thagard 1989). We have already discussed candidate inferences; we will now discuss each of the others in turn.

3.1 Schema Abstraction

One important kind of representational change is schema abstraction, which occurs when a common system derived from an analogy is highlighted, thereby increasing the possibility that it will be used again later (Gick and Holyoak 1983; Lowenstein et al. 1999). There are several lines of evidence that comparing structurally similar problems can lead to schema abstraction: (1) such comparison leads to improved performance on further parallel problems and promotes transfer from concrete comparisons to abstract analogies (as in the Loewenstein and Gentner (2001) developmental study discussed earlier; (2) several studies have shown that when participants write the commonalities resulting from an analogical comparison, the quality of their relational schema predicts the degree of transfer to another example with the same structure (e.g., Gentner et al. 2003; Gick and Holyoak 1983; Lowenstein et al. 1999).

Through schema abstraction, analogy can promote the formation of new relational categories (Gentner 2005) and abstract rules (Gentner and Medina 1998). One way this can occur is via *progressive alignment* – repeated schema abstraction across a series of exemplars. In this way, initially concrete, dimensionally specific representations are rendered more abstract by comparison and alignment. This kind of learning may be especially important in very young children. The idea is that close literal matches are easy for young children to perceive, because they are, in a sense, automatically aligned. This alignment results in a slight highlighting of the common relational structure, which can then seed further alignments with more distant examples.

A particularly dramatic example of early learning was found by Marcus et al. (1999), who found that through repeated exposure to relationally similar exemplars, infants can learn to recognize regularities in simple language-like stimuli. For example, if the infants had heard several instances of an ABA pattern, they would notice the shift to a novel (ABB) pattern. Kuehne et al. (Kuehne et al. 2000b) simulated this "infant rule-learning" using a model of learning by progressive alignment. This model, called SEQL (Kuehne et al. 2000a), forms abstractions across a set of exemplars by making successive structural comparisons (using SME) among exemplars. When a new exemplar is introduced, it is compared to the existing abstractions and (if sufficiently similar) assimilated into that abstraction, typically resulting in a slightly more abstract generalization. Exemplars that cannot be assimilated into any existing category (because they are too dissimilar from the existing generalizations) are maintained as separate exemplars.

The SEQL simulation was able to learn the language-like patterns within the same number of trials as the infants, and without pretraining (in contrast to

connectionist simulations of the same phenomenon, which required extensive pretraining (e.g. 50,000 trials; Seidenberg and Elman 1999). For example, when presented with new strings it found those with the same structure far more similar than those with different structure. Interestingly, although the simulation matched the infant data beautifully, its generalization was not a fully abstract rule. Rather, the generalization retained some surface features; yet because of the structural character of the matching process, SEQL still found new instances with matching structure to be much more similar than those with a different structure. These findings raise the tantalizing possibility that some of the seemingly abstract rules of grammar and logic may in fact be simply near-abstractions resulting from progressive alignment.

3.2 Re-representation

The third way that representations can be altered is through re-representation of the relations to create a better match between the two analogs (see Holyoak et al. 1994; Keane 1996; Kotovsky and Gentner 1996; Yan et al. 2003). For example, when people are given the analogy below, they typically arrive at the commonality "Each *got rid of* something they no longer wanted."

Walcorp divested itself of Acme Tires.

Likewise, Martha divorced George.

The re-representation of relations can occur in conceptual analogies like the above, but it can also occur in perceptual analogies. For example, Kotovsky and Gentner (1996) gave 4-year-old children a similarity task in which they saw simple three-shape patterns like those shown in Fig. 2. When given triads that showed the same relational pattern – e.g., symmetry – across different dimensions (as in the right triad in Fig. 2), children had great difficulty recognizing the similar pattern; they chose randomly between the two alternatives. However, when children were first asked about triads that varied on the *same* dimension (e.g., squares and circles that varied on the size dimension), they were then more able to subsequently recognize the pattern cross-dimensionally. These results suggest that this method of

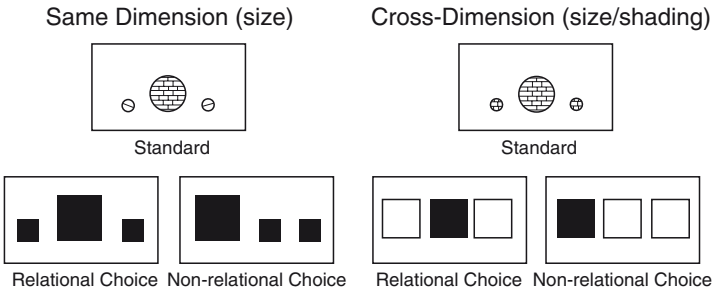

Fig. 2 Sample stimuli from Kotovsky and Gentner (1996)

progressive alignment – where highly similar items are compared first, followed by less similar items – fosters re-representation of the relevant relations.

Such re-representations could of course be temporary, in service of a particular task, but it seems likely that some re-representations can be learned and retained. Pervasive metaphors, such as "happy is up" (e.g., "After days of depression, his spirits finally lifted") (Lakoff and Johnson 1980) permeate natural language to an extent that suggests that at least some re-representations may become a more permanent part of our cognitive repertoire.

4 Analogical Retrieval

So far our focus has been on analogical mapping once the base and target have been mentally juxtaposed. However, explaining the use of analogy and similarity in reasoning requires some account of how potential analogs are accessed in long-term memory. Relational retrieval can be said to be the Achilles' heel of our relational capacity. There is considerable evidence that similarity-based retrieval, unlike the mapping process, is more influenced by surface similarity than structural similarity. Strong surface similarity and content effects seem to dominate remindings and to limit the transfer of learning across domains (Gentner et al. 1993; Holyoak and Koh 1987; Keane 1988; Novick 1988a,b; Reed 1987; Ross 1984, 1987, 1989).

In Gick and Holyoak's (1980, 1983) classic studies, participants often failed to access potentially useful analogs. For example, in one experiment (1980, E5), the rate of successfully solving a very difficult problem quadrupled (to 41%, from a baseline of 10%) for participants who were given an analogous story prior to the problem; but even so, the majority of participants failed to benefit from the analogy. However, when nonsolvers where given a hint to think about the story they had heard, the solution rate nearly doubled again to 76%. Because no new information was given about the story, it can be concluded that the analog was available in memory, but was not spontaneously retrieved. The structural similarity between the story and the problem was sufficient to carry out the mapping when both analogs were present in working memory, but not sufficient to produce spontaneous retrieval.

To test the functional distinction between kinds of similarity, Gentner et al. (1993) gave participants a large set of stories to remember and then later provided new stories that varied in their surface and relational similarity to the originals. Participants were asked to write out any original stories they were reminded of – the remindings that resulted were strongly governed by surface commonalities such as similar characters. However, as shown in Fig. 3, when asked to rate the similarity and inferential soundness of pairs of stories, the same participants relied primarily on higher-order relational commonalities, such as matching causal structure. Participants even rated their own surface-similar remindings as poor matches. This dissociation is also found in problem-solving tasks: remindings of prior problems are strongly influenced by surface similarity, but structural similarity better predicts success in solving the problem (e.g., Ross 1987).

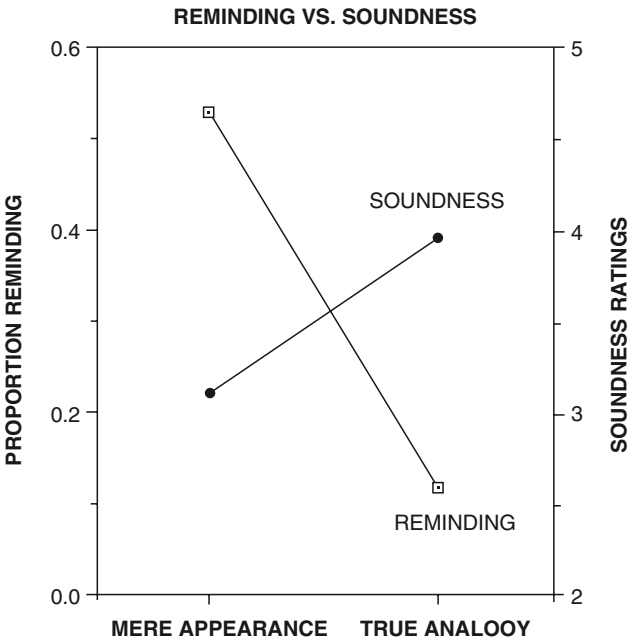

Fig. 3 Results from Gentner et al. (2003) showing that mere appearance matches produced more remindings, whereas true analogies were given higher soundness ratings

Overall, these are rather gloomy findings. Our poor capacity capacity for relational retrieval seems to belie our vaunted human ability for relational cognition. Yet, perhaps paradoxically, one remedy for poor relational retrieval is to make greater use of analogy during online learning and reasoning (e.g., Gick and Holyoak 1983). Studies by Loewenstein et al. (1999) and Gentner et al. (2003), for example, have shown that comparing analogous cases instantiating a complicated negotiation principle greatly improves transfer, such that those who were encouraged to compare the cases were more likely to apply the principle in a face-to-face negotiation task (in which it was appropriate) than were those who studied the cases without comparing.

Furthermore, these researchers (Gentner et al. in press) suggest that alignment-induced re-representation can even improve access to representations stored *prior* to the alignment. Whereas the above studies have shown that comparison during encoding facilitates future relational transfer, Gentner et al.'s recent work has shown that comparison at a later time can facilitate retrieval of material previously stored. Gentner et al. gave participants two cases instantiating a certain negotiation principle, then asked them to recall prior cases of the same principle. Those who were encouraged to compare the training cases were more likely to retrieve matching prior cases than those who read the training cases individually. This finding suggests that analogical encoding can provide a potent means of accessing our vast stores of relational knowledge.

5 Concluding Remarks

As an account of similarity and comparison, the alignment-based approach contrasts sharply with the featural and geometric (or distance) models, such as Tversky's (1977) contrast model and Shepard's (1962) multi-dimensional scaling model. Those models are concerned with the matching of features, with little or no attention to the relations among such features, and thus have difficulty coping with structured representations (see Goldstone et al. (2009), for a detailed discussion).

The alignment-based approach, in contrast, gives due priority to finding common relational structure. Structural alignment depends crucially on the relations among the entities being compared. It highlights the common relational structure, which in turn leads to re-representation and abstraction. Guided by systematicity, alignment also engenders new inferences – a key to generating knowledge.

Analogical processes are at the core of relational thinking, a crucial ability that, we suggest, is key to human cognitive prowess and separates us from other intelligent creatures. Our capacity for analogy ensures that every new encounter offers not only its own kernel of knowledge, but a potentially vast set of insights resulting from parallels in the past and future.

Acknowledgements This research was funded by ONR Grant N00014-08-1-0040. Correspondence concerning this chapter should be addressed to gentner@northwestern.edu or Dedre Gentner, Psychology Department, Northwestern University, Evanston, IL 60208.

References

Bowdle B, Gentner D (1997) Informativity and asymmetry in comparisons. Cogn Psychol 34(3):244–286

Clement J (1988) Observed methods for generating analogies in scientific problem solving. Cogn Sci 12(4):563–586

Clement CA, Gentner D (1991) Systematicity as a selection constraint in analogical mapping. Cogn Sci 15(1):89–132

DeLoache JS (1987) Rapid change in the symbolic functioning of very young children. Science 238:1556–1557

Falkenhainer B, Forbus KD, Gentner D (1989) The structure-mapping engine: algorithm and examples. Artif Intell 41:1–63

Forbus KD, Gentner D, Law K (1995) MAC/FAC: a model of similarity-based retrieval. Cogn Sci 19:141–205

Forbus K, Gentner D, Everett J, Wu M (1997) Towards a computational model of evaluating and using analogical inferences. In: Proceedings of the 19th Annual Conference of the Cognitive Science Society. LEA, Inc., NJ, London, pp 229–234

Gentner D (1983) Structure-mapping: a theoretical framework for analogy. Cogn Sci 7:155–70

Gentner D (1989) The mechanisms of analogical learning. In S. Vosniadou, A. Ortony (eds.), Similarity and analogical reasoning. London: Cambridge University Press, pp 199–241 (Reprinted in knowledge acquisition and learning, 1993, 673–694).

Gentner D (2003) Why we're so smart. In: Gentner D, Goldin-Meadow S (eds) Language in mind: advances in the study of language and cognition. MIT Press, Cambridge, MA, pp 195–236

Gentner D (2005) The development of relational category knowledge. In: Gershkoff-Stowe L, Rakison DH (eds) Building object categories in developmental time. Erlbaum, Hillsdale, NJ, pp 245–275

Gentner D, Christie S (2008) Relational language supports relational cognition in humans and apes. Behav Brain Sci 31:136–137

Gentner D, Markman AB (1995) Similarity is like analogy: structural alignment in comparison. In: Cacciari C (ed) Similarity in language, thought and perception. Brepols, Brussels, pp 111–147

Gentner D, Markman AB (1997) Structure mapping in analogy and similarity. Am Psychol 52:45–56

Gentner D, Medina J (1998) Similarity and the development of rules. Cognition 65:263–297

Gentner D, Rattermann MJ (1991) Language and the career of similarity. In: Gelman SA, Brynes JP (eds) Perspectives on thought and language: Interrelations in development. Cambridge University Press, London, pp 225–277

Gentner D, Toupin C (1986) Systematicity and surface similarity in the development of analogy. Cogn Sci 10(3):277–300

Gentner D, Wolff P (1997) Alignment in the processing of metaphor. J Mem Lang 37:331–355

Gentner D, Rattermann MJ, Forbus KD (1993) The roles of similarity in transfer: Separating retrievability from inferential soundness. Cogn Psychol 25:524–575

Gentner D, Loewenstein J, Thompson L (2003) Learning and transfer: A general role for analogical encoding. J Educ Psychol 95(2):393–408

Gentner D. Loewenstein J, Thompson L, Forbus K (in press) Reviving inert Knowledge: Analogical abstraction supports relational retrieval of past events. Cogn Sci

Gick ML, Holyoak KJ (1980) Analogical problem solving. Cogn Psychol 12:306–355

Gick ML, Holyoak KJ (1983) Schema induction and analogical transfer. Cogn Psychol 15:1–38

Glucksberg S, McGlone MS, Manfredi DA (1997) Metaphor comprehension: how metaphors create new categories. In: Ward T, Smith S, Vaid J (eds) Creative thought: an investigation of conceptual structures and processes. APA, Washington, DC

Goldstone, Day and Son, (2009) Comparison. -- **this volume**

Goldstone RL, Medin DL, Gentner D (1991) Relational similarity and the non-independence of features in similarity judgments. Cogn Psychol 23:222–264

Greiner R (1988) Learning by understanding analogies. Artif Intell 35:81–125

Heit E, Rubinstein J (1994) Similarity and property effects in inductive reasoning. J Exp Psychol Learn Mem Cogn 20:411–422

Holyoak KJ, Koh K (1987) Surface and structural similarity in analogical transfer. Mem Cogn 15:323–340

Holyoak KJ, Thagard P (1989) Analogical mapping by constraint satisfaction. Cogn Sci 13:295–355

Holyoak KJ, Novick LR, Melz ER (1994) Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In: Holyoak KJ, Barnden JA (eds) Advances in connectionist and neural computation theory, Analogical connections, vol 2. Norwood, NJ Ablex, pp. 113–180

Hummel JE, Holyoak KJ (1997) Distributed representations of structure: A theory of analogical access and mapping. Psychol Rev 104:427–466

Keane MT (1988) Analogical problem solving. Chichester, W. Sussex, England: E. Horwood. Halsted Press, New York

Keane MT (1996) On adaptation in analogy: tests of pragmatic importance and adaptability in analogical problem solving. Q J Exp Psychol 49/A(4):1062–1085

Kotovsky L, Gentner D (1996) Comparison and categorization in the development of relational similarity. Child Dev 67:2797–2822

Kuehne SE, Forbus KD, Gentner D, Quinn B (2000a) SEQL – Category learning as progressive abstraction using structure mapping. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society. Philadelphia, PA, pp 770–775

Kuehne SE, Gentner D Forbus KD (2000b) Modeling infant learning via symbolic structural alignment. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society. Philadelphia, PA, pp 286–291

Kurtz KJ, Gentner D, Gunn V (1999) Reasoning. In: Rumelhart DE, Bly BM (eds) Cognitive science: handbook of perception and cognition, 2nd edn. Academic Press, San Diego, pp 145–200

Lakoff G, Johnson M (1980) Metaphors we live by. University of Chicago Press, Chicago

Loewenstein J, Gentner D (2001) Spatial mapping in preschoolers: close comparisons facilitate far mappings. J Cogn Dev 2(2):189–219

Marcus GF, Vijayan S, Bandi Rao S, Vishton PM (1999) Rule-learning in seven-month-old infants. Science 283:77–80

Markman AB (1997) Constraints on analogical inference. Cogn Sci 21(4):373–418

Markman AB (1999) Knowledge representation. Lawrence Erlbaum Associates, Mahwah, NJ

Markman AB, Gentner D (1993a) All differences are not created equal: a structural alignment view of similarity. In: Proceedings of the 15th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Boulder, CO, pp 682–686

Markman AB, Gentner D (1993b) Splitting the differences: a structural alignment view of similarity. J Mem Lang 32:517–535

Markman AB, Gentner D (1993c) Structural alignment during similarity comparisons. Cogn Psychol 25:431–467

Medin DL, Goldstone RL, Gentner D (1993) Respects for similarity. Psychol Rev 100(2):254–278

Novick LR (1988a) Analogical transfer: processes and individual differences. In: Helman DH (ed) Analogical reasoning: perspectives of artificial intelligence, cognitive science, and philosophy. Kluwer, Dordrecht, The Netherlands, pp 125–145

Novick LR (1988b) Analogical transfer, problem similarity, and expertise. J Exp Psychol Learn Mem Cogn 14:510–520

Ortony A (1979) Beyond literal similarity. Psychol Rev 86:161–180

Paik JH, Mix KS (2006) Preschoolers' similarity judgments: Taking context into account. J Exp Child Psychol 95:194–214

Penn DC, Holyoak KJ, Povinelli DJ (2008) Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Brain Behav Sci 31:109–178

Reed SK (1987) A structure-mapping model for word problems. J Exp Psychol Learn Mem Cogn 13:124–139

Ross BH (1984) Remindings and their effects in learning a cognitive skill. Cogn Psychol 16:371–416

Ross BH (1987) This is like that: the use of earlier problems and the separation of similarity effects. J Exp Psychol Learn Mem Cogn 13(4):629–639

Ross BH (1989) Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. J Exp Psycho Learn Mem Cogn 15(3):456–468

Seidenberg MS, Elman J (1999) Do infants learn grammar with algebra or statistics? Lett Sci 284:434–436

Shepard RN (1962) The analysis of proximities: multidimensional scaling with an unknown distance function, I. Psychmetrika 27(2):125–140

Tversky A (1977) Features of similarity. Psychol Rev 84(4):327–352

Wolff P, Gentner D (2000) Evidence for role-neutral initial processing of metaphors. J Exp Psychol Learn Mem Cogn 26:529–541

Wolff P, Gentner, D. (under review) Structure-mapping in metaphor: evidence for a multi-stage model of metaphor processing. Cogn Sci

Yan J, Forbus K, Gentner, D (2003) A theory of rerepresentation in analogical matching. In R. Alterman & D. Kirsh (eds) Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society, Boston MA, pp 1265–1270

A Gestalt Perspective on the Psychology of Thinking

Michael Wertheimer

Abstract Gestalt theory, one of the major "schools" of psychology during the first half of the twentieth century, recently returned to prominence because of the enormous relevance to current research in cognitive science and other areas. Core concepts in Gestalt theory are dynamic self-distribution, structure, relational determination, organization, Prägnanz, reorganization, insight, and understanding. The most basic principle of Gestalt theory is that most wholes in nature are not merely the sums of their constituent elements, nor just more than the sums of their parts, but qualitatively entirely different from some additive product. Gestalten are dynamic structures the qualities and nature of which determine the place, role, and function of their constituent parts. Several examples illustrate how productive human thinking involves transforming a confused, opaque, incomprehensible problem situation into a clear, clean Gestalt or organization which makes sense, is coherent, and generates insight about the genuine nature of the problem structure and its solution.

1 Introduction

Gestalt theory, one of several "schools" of psychology that flourished during the first half of the twentieth century, recently became prominent again because of its relevance to current research issues in fields as diverse as cognitive neuroscience, perception, visual neuroscience, the psychology of art, social psychology, the study of personality – and problem solving and thinking. Indeed its origin and primary focus throughout its history has been the psychology of thinking. Gestalt psychologists' early theoretical formulations and empirical research studies also concentrated on the organization of perception, but its main concern has consistently been the cognitive processes involved in productive thinking. The two themes were

M. Wertheimer
Department of Psychology, University of Colorado, Muenzinger Psychology Building, Campus Box 345, Boulder, 80309, CO, USA
e-mail: Michael.Wertheimer@colorado.edu

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_4,

closely related, because successful problem solution, according to the Gestalt theorists, involves developing a clear perception or overview of the critical features of the problem situation, a process that typically requires an insightful reorganization of how the problem is viewed. Crucial in productive thinking is grasping the core or essence of the problem, understanding its key features, developing insight into its genuine nature, and not being distracted by irrelevant or superficial characteristics. When such reorganization occurs, when the solution "clicks" for the thinker, when the nature of the problem has been fully grasped, there typically is a satisfying "Aha!" experience; a previously murky, confused conception of the problem situation is transformed into a clear, simple, often elegant recognition of the true organization of the problem's structure and its solution.

2 Some Basic Concepts in Gestalt Theory

Gestalt theory emerged as a movement in protest against what it viewed as the excessively atomistic, elementalistic, or "andsummative" views that prevailed in philosophy and psychology in the late nineteenth and early twentieth centuries (Wertheimer 1980). Nature, the Gestalt theories insisted, is not typically composed of arbitrarily connected hook-ups of mutually indifferent atoms or parts, but is more veridically viewed as made up of dynamically integrated wholes or Gestalten. The characteristics of such wholes are not merely the sum totals of the characteristics of the parts making up the whole. Rather, conversely, the nature of the whole determines the nature of its parts – indeed determines the place, role and function of each part in the whole. The whole is not simply the sum of its parts, nor is the whole merely more than the sum of its parts; wholes are fundamentally entirely different from a bare sum total of their parts.

The most basic concept of Gestalt theory is, of course, *Gestalt* itself. The word has become part of the international technical language because the German term is difficult to translate into English and other languages. Rough equivalents in English are configuration, structure, form, shape, or pattern, but none of these fully captures the dynamic nature of the German word. An integrated, articulated whole, a Gestalt is an organized totality within which the nature, place, role and function of each part is precisely what it must be, given the nature of the whole. The parts are in dynamic interaction with each other and with the whole; they are not a mere bundle or concatenation of items that happen to be arbitrarily glued or hooked together by accidental proximity in space and time. A soap bubble is a good example of a Gestalt. *Dynamic self-distribution* of the soapy film assures that the thickness of the film is relatively uniform throughout the entire structure. Furthermore, the parts of the film are not indifferent to one another, but are intimately interrelated; if you prick the bubble with a pin, the local disturbance has devastating consequences for the entire *structure*. The dynamic interrelationships among the parts of a whole are determined by the nature of the entire structure itself; the parts of a Gestalt are not constrained nor arbitrarily held together by external mechanical forces. A related concept is

relational determination: the relationship of parts to each other and to the whole determines the nature of each part and its place, role and function in the whole. The same "part" or "element" may play a very different role in different wholes. Thus the note of C played on a piano can have a striving quality that requires resolution if it happens to be part of a D7 chord (resolved by a tonic G chord in classical harmony), but be a satisfying closing tone for a melody played in the key of C (or serve as a plaintive diminished third in an A minor sequence).

Another basic concept in Gestalt theory is *organization.* Gestalten typically display *Prägnanz,* a characteristic inherent organization that is as "good" as the prevailing conditions allow. This principle of Prägnanz applies to all Gestalten: perceptual, cognitive, social, physiological, psychological, and physical. Examples include electrical or magnetic fields, the human body, a soap bubble, successful works of art, fine musical compositions, or functioning social practices. Parts of the whole are far from indifferent to one another. In perception, for example, what gets organized into a single unit is determined by the qualities of the parts: their similarity, their proximity, whether together they generate a closed figure, whether through time they are undergoing a common fate (such as moving in the same direction); these "principles of perceptual organization" are still discussed in almost every introductory psychology textbook.

The concept of *reorganization*, while it also applies to perception, is crucial to the understanding of human thought. The typical sequence in problem solving involves going from a lack of understanding, from a state of frustrated confusing vagueness about what is going on and what is required to solve the problem to a state of *insight* or the achievement of a veridical impression of the state of affairs. This reorganization or "Aha!" experience results in a view of the genuine nature of the problem situation, understanding how everything fits together – and a clear conception of what is required to solve the problem. Achieving *understanding* means going from a situation that appears senseless, opaque, and incomprehensible to one in which central features of the problem are recognized, irrelevant aspects are ignored, and a meaningful solution is apparent. The reorganized conception of the problem situation does justice to the *radix* (or root, or essence) of the problem. If the task, for example, is to construct a toy bridge out of wooden blocks, then the color of the blocks is irrelevant to whether or not the bridge will stand, but there is a crucial relationship between the distance separating the uprights and the length of the cross piece.

Reorganization is a common process in perception, as when an initially confusing visual stimulus is suddenly recognized for what it actually is (a well-camouflaged animal in its natural habitat – say a speckled flounder lying in shallow water on sand – may be functionally invisible while stationary, but can be readily identified if it moves). Many ambiguous figures can be seen in different ways, some of which are usually more compelling than others.

Consider the two patterns in Fig. 1. What are they?

The left one could be seen as a crude profile of a smiling face with a squinting eye.

The right one might be a ghostly figure looking over a fence, perhaps with the ears of a dog sticking up to its left. These are reasonable ways to see the patterns.

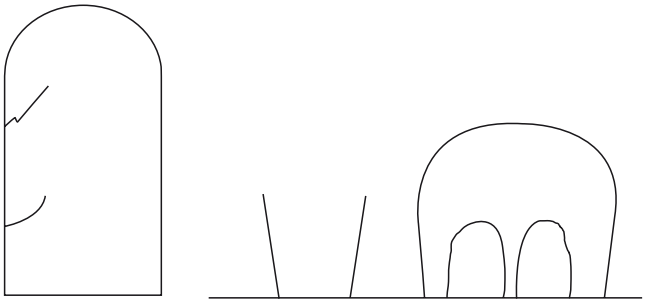

Fig. 1 Two ambiguous patterns

But there are other ways to organize them. On the left: the frame is a doorway, the "smile" the tail of a dog, and the "squinting eye" the fixed bayonet at the end of a rifle. A soldier and his dog are walking past an open portal. For the other pattern: late in the evening in an office building, the floor is being washed by hand. Now the "dog ears" become the sides of a cleaning bucket and the "ghost's eyes" become the heels of a custodian who is viewed from a somewhat compromising angle, with what had been the top of the ghost's head becoming the janitor's rump. While these reorganizations are modest, they do yield completely different perspectives.

Proper organization or reorganization can result in understanding, "catching on," insight, and the achievement of meaning. And it is crucial, of course, in successful human thinking.

3 Some Historical Background on Gestalt Theory

Holistic thought has existed for millennia, including in some ancient Greek philosophies. But the success of elementaristic, mechanistic thinking in the physical sciences was accompanied by similar theorizing in psychology and philosophy over the centuries, and was still the prevailing view well into the nineteenth century. In the middle of the nineteenth century, John Stuart Mill, in opposition to his father's influential mechanistic associationist theory of mind, proposed a kind of "mental chemistry": mental compounds, rather than being simply the sum-total of their associated constituent mental elements (sensations or ideas) can, like chemical compounds, have emergent properties that are radically different from those of the "elements" of which they are composed (contrast the liquidity of water with the properties of the two gaseous elements, hydrogen and oxygen, of which water is composed). Yet late in the nineteenth century, mental wholes were still widely viewed as basically nothing more than the sum of their mental parts. In a paper in 1890, Christian von Ehrenfels argued that most mental wholes are *more than* the mere sums of their parts. Mental wholes are the sums of their elements plus one more element: a Gestalt quality. Thus a square is the sum of four equal straight lines

plus four right angles plus the element of "squareness." A melody is the sum of the notes that are in it plus the particular Gestalt quality of that melody; you can change all the "elements," the notes, by transposing the melody to a different key, but since the added element of the "form quality" is still the same, you can still readily recognize the melody. A distinctive feature of these Gestalt qualities that are added to the other elements, von Ehrenfels argued, is indeed their transposability to different elements (different notes, lines of different length or color, etc.).

A transition had thus occurred from viewing wholes as merely the sum of their parts to viewing them as *more than* the sum of their parts. But the Gestalt theorists early in the twentieth century proposed a radically new view: The entire additive view is wrong. Most wholes are epistemologically and psychologically prior to their parts, and are totally different from a sum of their parts or even the sum of their parts plus some other kind of "element." Organization does not occur as it were "from below up," adding things together, but "from above down," since the nature of the whole determines the nature of its parts. Indeed the parts do not exist as parts until there is a whole within which they function as meaningful parts.

Gestalt theory in psychology was founded in Germany by Max Wertheimer, Wolgang Köhler, and Kurt Koffka. Its work was devoted to the analysis of the structure of music (Wertheimer 1910), the creative numerical thinking of indigenous people (Wertheimer 1912a), productive thinking (Wertheimer 1920, 1945), perception (Wertheimer 1912b, 1923), broad theoretical, systematic statements (Koffka 1922, 1935; Köhler 1920, 1929, 1969; Wertheimer 1922), and eventually philosophical issues such as the nature of truth, ethics, democracy, and freedom (Wertheimer 1934, 1935, 1937, 1940). Students of the original Gestalt theorists, mostly in the United States after the original Gestalt theorists emigrated there in the 1930s for political reasons, applied the Gestalt approach to art (Arnheim, 1954), social psychology (Asch, 1952), problem solving (Luchins, 1942; Duncker 1945), learning (Katona, 1940), personality and social processes (Lewin, 1935, 1948) and other areas.

While the prominence of Gestalt theory waned by the middle of the twentieth century as the "age of schools" generally waned as well in psychology, it again became significant late in the twentieth century as its relevance to then-current research issues became clear once more (King and Wertheimer 2005; Spillmann 1999, 2001). That Gestalt theory was more influential than most other holistically oriented approaches during the twentieth century can probably be attributed to its emphasis upon precise formulations and rigorous experimental methods (Ash 1995; King and Wertheimer 2005).

4 The Gestalt Approach to Thinking

Much Gestalt writing about thinking concentrated on attempts to do detailed justice to instances of successful problem solving. Since the nature of every problem situation is specific to that problem itself, an analysis of creative and insightful

solutions to problems must concentrate on the conditions prevailing in a specific problem situation. Wertheimer's (1912a) paper on the numerical thinking of indigenous peoples is replete with specific structural examples: how many sticks of which length, strength, and thickness are required to frame a hut; if you break a wooden spear in half you don't have "two" in any meaningful sense but a broken spear or some useless pieces of wood (one of which might have a potentially reusable spear point); if you add two horses you have two horses, and if you add two people you have two people – but if you add two horses and two people you may have two riders; "much," "a few," "some," etc. may be more useful quantifiers than specific numerals (think of rice and other commodities); and if you cut a chain of eight links in half, you have two chains of four links each; if these in turn are cut in half you have four chains of two links each (hardly chains any more); and if you continue cutting each "part" in half, you no longer have any semblance of a chain but only eight rings – and if you continue the cutting, you don't even have links any more, but just 16 C-shaped objects.

Köhler (1917, 1925) published an influential book based on experiments he performed on problem solving by captive chimpanzees while he was marooned on the island of Tenerife during World War I. In order to obtain a banana suspended higher in their cage than they could reach even by jumping, several chimps constructed towers of sturdy boxes that were strewn about the cage. While sometimes the towers of two or more boxes were precarious, standing only long enough for the chimps to climb up them and snatch the banana, they were clearly built in the right place (directly under the banana, and not at some senseless other place in the cage far from the banana). Köhler reported that one chimp even took Köhler by the hand and steered him underneath the suspended banana – and before Köhler realized what was going on, had used Köhler as a ladder to retrieve the banana. More mundane solutions to obtaining otherwise unreachable lures included breaking a small branch off a dead tree in the enclosure to use as a rake to pull in the fruit, and one bright ape even inserted one short piece of bamboo into another short one so as to make a device long enough to retrieve a banana outside the cage which could not be reached with either stick alone. There was a clear relationship between what the chimps did and the structural requirements of the problem situation, and some of their solutions were quite ingenious.

Many mathematical examples exist for demonstrating the phenomenon of insight. Consider the following problem. As in algebra, let different letters stand for various numbers, so that the sequence "abc,abc" might be 342,342 or 497,497 or 213,213, etc. Any integers can be substituted for the individual letters, as long as the substituting remains consistent. Now: your task is to prove that any six-digit number with this abc,abc structure is divisible by 7 without remainder. If you have not encountered this problem before, ponder it a while before continuing to read on. The typical person who has not seen this problem before tends to be stumped. It may be true that 764,764 and 918,918 and 546,546, etc., are indeed divisible by 7 without remainder, but the problem seems to lack a "handle." What features of the information given provide a clue that could be used in *proving* that *every* number of this form has this same property? A first hint is to ask, by what is it divisible

without remainder? One, of course, since if you multiply abc,abc by one, you get abc,abc, and abc,abc itself, which times one is abc,abc. Any others? Try to divide abc,abc by abc: what is the quotient? It turns out to be 1,001; 1,001 times abc equals abc,abc. Have you made any progress? Yes: if abc,abc is divisible by 1,001 and 1,001 is divisible by 7, then abc,abc must be divisible by 7 without remainder! You try it, and discover that 7 goes into 1,001 exactly 143 times. So there you are! Yes, every number of the form abc,abc is divisible by 7 without remainder, because abc,abc is divisible by 1,001 without remainder; and 1,001 is divisible by 7 without remainder. But that is not all. You now can prove that every number of the form abc,abc is also divisible without remainder by 11 – and by 13 – as well as by 7. How? (The procedure is of course structurally the same; you show that like 7 and 143, 11 and 13 are factors of 1,001, since both are factors of 143.)

Reorganization and restructuring are not, of course, limited to perception or mathematical puzzles. One story the Gestalt theorists sometimes used in their lectures involves a caravan crossing a desert and approaching an oasis. The caravan consists of many camels and other animals belonging to a wealthy man who is mounted on a horse. He is accompanied by two lieutenants also riding on horses – and an old wise man bringing up the rear of the caravan on foot. As the oasis comes into view the wealthy man proposes to his two assistants: "To that one of you whose horse reaches the oasis *last*, I will donate this donkey laden with gold." Delighted, the lieutenants continue to ride towards the oasis. But soon one slows down a bit – and then so does the other, each waiting for the other to get ahead of him. Before long they are both stopped, and the caravan passes them by. When the man bringing up the rear comes upon them, they have dismounted and are sitting in the shadows of their horses, waiting for the other one to become so hot and thirsty that he would get back up on his horse and ride to the oasis despite the wager. "Why are you sitting on the hot sand this close to the oasis, rather than going on to get some water and real shade?" asks the wise man. The lieutenants tell him of their employer's generous offer. "Would you like me to give you some advice?" asks the wise man. Desperate, the lieutenants request his counsel. He says two words to them, whereupon they get up, jump on the horses, and race towards the oasis. Why? What did the wise man say to them? Once again, if you have not run across this puzzle before, think a bit before reading further. Why would the lieutenants remount the horses and race toward the oasis if that lieutenant whose horse gets to the oasis *last* would be the winner? The lieutenants' behavior seems to make no sense, given the conditions of the wager. But reread the wealthy man's offer carefully once again. He did not offer the prize to the *man* who got to the oasis last, but to the lieutenant whose *horse* got to the oasis last. Does that make a difference? Yes, if lieutenant A rode *lieutenant B's horse* and got lieutenant B's horse to the oasis before lieutenant B got lieutenant A's horse there, then lieutenant A would win; and the same logic, of course, applies to lieutenant B. Hence what the wise man said to the lieutenants is, "Trade horses," which they did, with each racing to get the other one's horse, upon which he was riding, to the oasis before the other one got his horse there. Now the action of the two lieutenants ("jumping on the horses and racing to the oasis") makes sense.

Productive thinking applies to many other situations as well. One example (Wertheimer 1945) involves two boys playing badminton. The older boy is much better at the game than the younger, and, not surprisingly, the younger one decides he does not want to play any more, because he loses all the time. The older one enjoys hitting the shuttlecock expertly over the net so much that he tries to cajole the "spoilsport" into playing some more, but without success. How can the game be restructured in such a way as to solve the problem, so that the older player can continue to enjoy his skill and the younger one is not frustrated because he loses all the time? One solution involves, again, a reorganization, a restructuring. Rather than trying to win points by hitting the shuttlecock in such a way that it lands in bounds but is hard for the opponent to return, the aim should be to make the shuttlecock easy to return: try to set a record for how often the shuttlecock can be sent back and forth across the net without hitting the ground. In this way, a "win-lose" situation is adroitly reorganized into a satisfying "win-win" situation.

The Gestalt approach was applied in analyses of an enormous variety of phenomena and examples of productive thinking, teasing out what the crucial root (again, "radix") features of each instance are, how reorganization or restructuring is required in order for clarity, insight, understanding – indeed a solution – to occur. There was also extensive work on the issue of transfer: If someone truly understands a problem class and its solution process, then it is far more likely that transfer of this skill will occur to problem situations that are superficially different but structurally similar to the initially learned problem class, than that transfer will occur if the initial learning was automatic, "blind," by rote memory, mechanized, and without insight or genuine understanding. Another area in which the Gestalt approach proved useful concerned "functional fixedness" (Duncker 1945), or the phenomenon that one solution strategy in a problem area that has worked successfully before may become so automatic as to preclude the discovery of a still more elegant and simpler solution to similar or even the same kind of problem, a phenomenon which has also been studied under the rubric "mechanization in problem solving" (Luchins 1942).

5 Concluding Thoughts on the Gestalt Approach to Thinking

Productive thinking involves creative transformation or reorganization of a situation that is confused, opaque, and incomprehensible into one which makes structural sense and that demonstrates understanding of and insight into the crucial features of the problem situation. The reorganized perspective does justice to the central aspects of the problem and recognizes irrelevant features for what they are. Genuine productive thinking is not a process of combining inert elements of a problem situation in the right way, but a process of achieving insight into the crucial, structural features of the problem. What is learned by insight is far better retained than something learned by rote memorization. Insight avoids stupid errors that ignore essential features of the problem situation. Genuine understanding can

be readily transferred to new situations that are structurally similar, even if the new situations may be superficially very different. Memorization or automatic, "blind" application of previously learned strategies is unlikely to lead to productive, insightful reorganization; each new problem must be approached with an open mind, in an attempt to discriminate what is crucial and what is trivial in the effort to achieve a structurally adequate perspective on the problem. Finally, achieving insight is often its own reward; genuinely understanding something after it was previously puzzling, murky, confusing, and frustrating is one of the most satisfying pleasures in human experience.

References

Arnheim R (1954) Art and visual perception. University of California Press, Berkeley, CA
Asch SE (1952) Social psychology. Prentice-Hall, New York
Ash MG (1995) Gestalt psychology in German culture, 1890–1967: holism and the quest for objectivity. Cambridge University Press, Cambridge, England
Duncker K (1945) On problem solving. In Psychological Monographs 58, Whole No. 270 (Original German edition 191)
Katona G (1940) Organizing and memorizing. Columbia University Press, New York
King DB, Wertheimer M (2005) Max Wertheimer and Gestalt theory. Transaction Publishers, New Brunswick, NJ
Koffka K (1922) Perception: an introduction to the Gestalt-Theorie. Psychological Bulletin 19:531–585
Koffka K (1935) Principles of Gestalt psychology. Harcourt, Brace, New York
Köhler W (1920) Die physischen Gestalten in Ruhe und im stationären Zustand. Vieweg, Braunschweig, Germany
Köhler W (1929) Gestalt psychology. Liveright, New York
Köhler W (1969) The task of Gestalt psychology. Princeton University Press, Princeton, NJ
Kölher W (1925) The mentality of apes. Harcourt, New York, Brace Original German edition, 1917
Lewin K (1935) A dynamic theory of personality. McGraw-Hill, New York
Lewin K (1948) Resolving social conflicts. Harper, New York
Luchins AS (1942) Mechanization in problem solving: the effect of *Einstellung.* In: Psychological Monographs 54: Whole No. 248
Spillmann L (1999) Gehirn und Gestalt. Psychologische Beiträge 41:458–493
Spillmann L (2001) Gehirn und Gestalt: II Neuronale Mechanismen. Kognitionswissenschaft 9:122–143
Von Ehrenfels C (1890) Über Gestaltqualiäten. Vierteljahresschrift für wissenschaftliche Philosophie 14:249–292
Wertheimer M (1910) Musik der Wedda. Sammelbände der internationalen Musikgesellschaft 11:300–309
Wertheimer M (1912a) Über das Denken der Naturvölker: I Zahlen und Zahlgebilde. Zeitschrift für Psychologie 60:321–378
Wertheimer M (1912b) Experimentelle Untersuchungen über das Sehen von Bewegung. Zeitschrift für Psychologie 61:161–265
Wertheimer M (1920) Über Schlussprozesse im produktiven Denken. De Gruyter, Berlin, Germany
Wertheimer M (1922) Untersuchungen zur Lehre von der Gestalt: I Prinzipielle Bemerkungen. Psychologische Forschung 1:47–58

Wertheimer M (1923) Untersuchungen zur Lehre von der Gestalt: II. Psychologische Forschung 4:301–350
Wertheimer M (1934) On truth. Soc Res 1:135–146
Wertheimer M (1935) Some problems in the theory of ethics. Soc Res 2:353–367
Wertheimer M (1937) On the concept of democracy. In: Ascoli M, Lehmann F (eds) Political and economic democracy. Norton, New York, pp 271–283
Wertheimer M (1940) A story of three days. In: Anshen RN (ed) Freedom: its meaning. Harcourt, Brace, New York, pp 555–569
Wertheimer M (1945) Productive thinking. Harper, New York Multiple translations, and enlarged edition 1959, 1982
Wertheimer M (1980) Gestalt theory of learning. In: Gazda GM, Corsini RJ (eds) Theories of learning: a comparative approach. Peacock, Itasca, IL, pp 208–251

Thought and Reality

A Philosophical Conjecture About Some Fundamental Features of Human Thinking

Albrecht von Müller

Abstract Trying to understand how thinking works cannot be separated from trying to understand how reality works. A recent approach to understanding how reality actually "takes place" postulates two complementary aspects. There is a "factual aspect of reality" which is characterized by well-defined predications, causal closure, and local spacetime. But there is a complementary, "statu-nascendi" aspect of reality which addresses how facts, and with them local spacetime, come into being in the first place. This aspect of reality is inherently constellatory, i.e. the constellations of components are the most basic phenomena – somewhat similar to Gestalt phenomena in the visual domain. Human thinking is interpreted as a highly advanced cognitive adaptation to this irreducible Janus-headedness of reality. In parts it can be well defined, in parts it just cannot – because it must leave room for the on-going self-unfolding of meaning. This self-unfolding of meaning is interpreted as the most accurate semantic approximation to the ongoing self-unfolding of reality. It is, thus, not a *bug* but a crucial *feature* of complex thinking. Unlike formal languages, which structurally correspond to the factual aspect of reality, natural language is capable of dealing with both aspects of reality: its facticity and its coming into being.

1 Introduction

Human thinking is probably the most complex process in the universe. But, very much like the phenomenon of time in which we are also comprehensively embedded in our existence, it is both extremely close to us, and also extremely difficult to grasp conceptually.

To strive for a better understanding of human thinking has at least four rather distinct motivations:

A. von Müller
Parmenides Foundation, Kirchplatz 1, D-82049 Munich/Pullach, Germany
e-mail: avm@parmenides-foundation.org

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_5, © Springer-Verlag Berlin Heidelberg 2010

- as a pinnacle of evolution it is a most fascinating issue for scientific scrutiny in itself,
- as a highly distinctive feature it sheds light on who we are, i.e. on the question "what makes us human?",
- as a competence it is quintessential for coping with the challenges of an increasingly complex world,
- as a production factor it becomes the key value generation process in a knowledge-based economy.

We are far away from understanding how human thinking actually works. A state-of-the-art answer to these questions must integrate insights from at least two rather heterogeneous approaches: the age-old philosophical quest for thinking and its relation to reality, and the more recent insights from cognitive neuroscience (including cognitive psychology, neuroinformatics and evolutionary anthropology). In the Parmenides Center for the Study of Thinking we are working towards such a "bipedal" theory of thinking.

The present paper introduces a philosophical conjecture about the relation between thinking and reality. The general approach is to interpret the human faculty of complex thinking as the most advanced adaptation of cognitive evolution to the way reality works.

It is a very good and successful tradition in science to try to address one issue at a time. I apologize for not following this commendable habit in this article, but there seems to be a compelling reason. I am convinced that complex thinking is – at least to a very large extent – a successful phylogenetic adaptation to the way reality actually works. (If thinking would consist predominantly of "confabulations" that have nothing to do with what is actually going on "out there", it would never have survived the pressures of evolutionary selection.) So, I see the conscious experience and cognition of human beings as the hitherto most advanced and most sophisticated way of coping with reality – as it actually is. If this assumption is right, it means that to understand how reality works is a prerequisite for understanding how thinking works – as the latter is, to a large extent, an evolutionary adaptation of the first.

And here the problem begins: At the most fundamental level we do not have a clue in modern science how reality works. Classical physics, and by this I mean Newtonian mechanics, Maxwell's electrodynamics and both, special and general relativity, is an extremely successful family of theories. But, eventually they all imply a comprehensively determined block universe in which nothing genuinely novel can happen, and in which the experience of time and especially the experience of a present are basically subjective confabulations, the "sticky illusion" Einstein was referring to.

On the other hand we have quantum physics. Its predictions have been proven right with even higher accuracy than the other three theories together. But there is no generally accepted interpretation of what quantum physics tells us about reality. Some opt for a quasi-classical interpretation, avoiding genuine novelty, but e.g. at the cost of an unimaginable inflation of the number of universes at every single moment in time (as each possibility gets realized its own universe). Although this position cannot be proven to be wrong *within* physics, I think it is highly questionable on philosophical grounds and, even more important, it is, at least for me,

"just too ugly to be true". The other camp believes that the future is open, i.e. that the reduction of state is a meaningful process that really takes place. This position, however, implies a full-fledged contradiction between general relativity and quantum physics – right at the most fundamental level of our understanding of reality.

Working since three decades on both, our understanding of time and reality on the one hand, and on the question, how thinking works, on the other, I come to the conclusion that we cannot really separate and isolate the two issues from one another: we will not be able to answer the second question without significant progress regarding the first – and vice versa.

For this reason the present article begins with a short summary of ongoing own work on a modified account of time and reality (see Literature). The main difference to the traditional view is that it assumes a complex notion of time in which the linear sequential character is just one of two aspects. The other, orthogonal, aspect is an expanded timespace of the present – in which the genuinely novel can occur. Only via this taking place of reality in the timespace of the present, facts materialize – and with them local spacetime starts to become applicable. In a nutshell, this approach claims two complementary aspects of reality: facticity on the one hand (successfully addressed by the family classical theories) and the "taking place of reality as such" on the other, addressed by quantum physics. Human cognition, and especially complex thinking, is seen as a successful adaptation to *both* aspects of reality and time – and understanding it requires, therefore, to start with looking at "how time and reality work".

Based on these foundational considerations about reality, one can then ask how thinking actually works. And there we see that both, the basic architecture of human thinking and many of its characteristic features become understandable as efficient and successful adaptations to this Janus-faced reality.

2 A Brief Summary of a Novel Account of Time and Reality

Time is until today probably the most mysterious of the fundamental concepts, both in philosophy and physics. Yet, de facto there is a broad implicit consensus. It holds that the primary feature of time is to provide – one way or the other – what we perceive as the sequential order of events. This is even common ground between relativity and quantum physics.

In the following a fundamentally different way to think about time is developed: The sequential structure is introduced only as a complementary, but in a way even derivative feature. The primary feature of time is to provide an expanded, but not yet sequentially structured temporal platform which constitutes the "stage" on which everything that is takes place. This stage is the "time-space of the present" – in which everything that is, comes into being. Only as facts materialize – on that stage, a prior starts to separate from a later and the sequential aspect of time starts to unfold. *The sequential notion of time applies, therefore, only to the factual aspect of reality.*

This notion of time differs profoundly from "presentism" according to which only the present exists, but in which this present is again conceptualized within the framework of a linear sequential time. Here, instead, both, the timespace of the

present and the sequential aspect of time (with past, point-like now and future) are complementary aspects of the taking place of reality. The present is no longer reduced to a point-like now; it is expanded, but this expandedness is "orthogonal" to the linear sequential aspect of time. Due to its orthogonality the present can now also play an objective role in physics, which is needed for understanding e.g. the state reduction as the occurrence of something genuinely novel (which is inconceivable in a purely classical or relativistic conceptual framework).

Only via this primordial "taking place" in the timespace of the present, what occurs becomes a fact – and only by doing so it gains its well-defined position in local spacetime. What occurs is "taking (its) place" in local spacetime in the most literal sense.

Natural laws assure the constitutive continuity of the factual aspect of reality. Their predictive power is a function of this – but also limited to it. In this way, the comprehensive determinacy of the factual aspect of reality, the "block universe" of relativity, becomes compatible with the objective indeterminacy i.e. the occurrence of something genuinely novel in the state reduction of quantum physics.

These are two complementary aspects of reality: Quantum physics addresses the taking place of reality – in the – not yet sequentially structured – time space of the present and this implies genuine novelty. Classical and relativistic physics, instead, are focused on the factual aspect of reality. Facticity and spacetime locality are functionally equivalent notions.

Neither of these two views allows for a comprehensive description of reality: "Measurement", the (asymptotic) transition form a coherent to a decoherent state implies – and requires – (asymptotic) facticity. On the other hand, the inevitability of singularities in general relativity shows that the fabric of local spacetime also cannot be thought of as an all-encompassing canvas – respectively that facticity cannot be the only aspect for describing reality.

In this new account of time and reality our human perception of a present is no longer pushed aside as a subject-side confabulation that has no role in physics (a fact that Einstein deplored explicitly in his discussion with Carnap). In evolutionary terms, such a costly, but dysfunctional distortion could not have survived the pressures of evolutionary selection. Instead, the experience of the present – inseparably linked to the phenomenon of consciousness – is seen as the hitherto most advanced form of higher cognition. Together with the (orthogonal) perception of the factual aspect of reality, it allows for a much richer and more accurate perception of the way in which reality actually takes place.

In this vain it is argued

- that the classical and relativistic physics deal primarily with the factual aspect of reality and that, therefore, the sequential notion of time is sufficient – at least to a large extent,
- that quantum physics, instead, addresses essentially how facts come into being in the first place – and, therefore, requires a much richer notion of time and reality,
- that the mathematical apparatus of quantum physics implicitly anticipates already much of both, a radically different, present-centered notion of time and the related, significantly richer notion of reality,

- that the discrepancy between classico-relativistic and quantum physics cannot be overcome by just quantizing gravity, but that we need to go back to the conceptual drawing board and to make their categorial foundations part of our physical theories.

All scientific theories are based on categorial foundations. Categories are the most basic patterns of thought that shape and constrain all we can think. Unlike hitherto assumed, these most basic structures of thought do not come in isolation but as categorial frameworks that are characterized by strong internal interdependencies. Classical and relativistic physics are based on a categorial framework that consists of four interdependent components:

- a *Boolean predication* space characterized by the principle of "tertium non datur",
- the *linear sequential notion of time* with a no-longer present past, a point-like now, and a not-yet present future,
- the *principle of causality* in the sense of causal closure according to which for anything that happens a sufficient cause exists,
- the *subject/object dichotomy* according to which there exists a clear-cut distinction between observer and observandum.

None of these four components can be dropped without destabilizing all others. They constitute an integral and coherent categorial apparatus. The common denominator of this categorial framework is comprehensive separability. This framework eventually implies a block universe and it applies to the factual aspect of reality – and only to it. Facts, however, are only the traces left behind by the taking place of reality. Quantum physics, instead, addresses also how facts come into being in the first place. For addressing also this prefactual or "statu-nascendi" aspect of reality a profoundly different, second categorial apparatus is needed, in addition. It consists again of four interrelated components that cover the same four functional slots; (a) the structure of the predication space, (b) a notion of time, (c) a pattern how events are linked, and (d) a basic epistemological setting.

In case of the second categorial apparatus these four constituents are

- a *paratactical predication space* allowing for constellations of propositions, but without the possibility of logical conclusions,
- the *timespace of the present* as an expanded, but not yet sequentially structured temporal platform on which reality takes place,
- the principle of *autogenetic unfolding* according to which something "becomes what it is", in the absence of external causation,
- the structure of *strong selfreferentiality* which appears whenever something refers to itself in its entirety.

Taken in isolation and projected against the rest of the first, the "classical" categorial framework, each of these four components leads immediately to inconsistencies. Taken together, however, they form a second, categorial apparatus in its own right – which is again inherently consistent. This second framework is complementary to the first and allows addressing the taking place of reality – as it actually occurs, i.e. in the time-space of the present. If something genuinely novel comes into being this cannot occur in the

past or the future. It can only occur in the present, and, therefore, genuine novelty implies an "objective role" for the present.

This "objective present", however, is not in contradiction with relativity. In the framework where it belongs, i.e. in the second categorial apparatus, it doesn't define a specific, point-like now that would be mandatory for the entire universe. Instead, as an "orthogonal" feature of time in its own right, its inseparable expandedness can "host" all possible now's – and the related paratactical predication assures that "contradicting" predications are just constellations of ascribing "a" and "non a" – coexisting in the same, expanded time-space. Due to the absence of formal conclusions in paratactic predication, these superposition states do not lead to the "ex-falso-quodlibet catastrophe" that they would imply in a Boolean predication space.

The four big "enigmas" of quantum physics, entanglement, superposition, uncertainty and objective indeterminacy, are the foot-print of the statu-nascendi aspect of reality – for which the second categorial apparatus is required. Only if we project this taking place of reality – erroneously – already against the categorial apparatus that belongs to the factual aspect of reality, these four features become "enigmas" and conceptually insurmountable problems.

Up to now one was not aware of the existence and the crucial role of these underlying categorial frameworks. Therefore, one tried to give up or modify elements of the first framework in isolation, e.g. by giving up causality for state reduction but continuing to work with linear sequential time as if nothing had happened. This resulted not only in the well-known "enigmas" of quantum physics, but it also created the hitherto unsolvable contradictions between quantum physics and general relativity.

These discrepancies, however, do not constitute fatal inconsistencies. They are the logical consequence of the phenomenon that classico-relativistic and quantum physics focus on different aspect of reality: the first on the factual aspect of reality, the latter on how facts come into being in the first place.

Only by gaining insight into the fundamental difference of these two complementary aspects of realty, and by applying the appropriate categorial framework, we will become able to overcome this essential rift of modern physics. This, however, requires that we dig still one layer deeper and to make their categorial underpinnings part of our physical theories.

3 Human Thinking as an Adaptation to a Janus-Headed Reality

If reality is inherently characterized by these two complementary aspects, facticity and statu-nascendi, human thinking – as the phylogenetically most advanced form of cognition – is likely to be structurally adapted and tuned to this "Janus-headedness" of reality.

There should be one set of cognitive processes that enable us to deal with the factual aspect of reality, and this are the well-defined or well-definable "ratiomorphic" operations. Instead, for the inherently self-referential and autogenetic aspect

of reality, i.e. reality in statu-nascendi, well-defined operations are structurally inadequate. In order to come grips with this – not epistemic, but ontic – incompleteness and uncertainty we have to look for profoundly different mechanisms, for which I use the notion "logic of constellations".

Sine Aristotle's great effort to define the rules of thinking "in abstracto" most of the history of "logic" focused on abstract rules for correct concluding. Heraclites, instead, had still a much richer notion of logic. For him "logos" was the overarching and all pervading principle according to which all of reality unfolds. Only Hegel formulated, more than 2000 years later, a similarly rich and encompassing notion of logic.

Based on what we saw regarding the Janus-headedness of reality, the reduction of logic to rules of formal correctness is definitively too narrow. It only covers the needs of the factual aspect of reality. The statu-nascendi aspect, instead, corresponds as we saw to a paratactical predication space. The "logic of constellations" describes what happens in a paratactical predication space, i.e. how *meaning unfolds* in constellations of semantically already meaningful components (e.g. words, concepts or propositions – but, in the case of art, also well beyond the domain of language).

Asking how this unfolding of meaning actually occurs, I propose to identify three closely interrelated, but nevertheless distinguishable dynamics.

- the first is the mutual interpretation of the components of the constellation, i.e. the "horizontal" dimension in the autogenetic unfolding of meaning.
- the second is an emergent "overarching meaning" of the entire constellation which constitutes, so to speak, "vertically" out of and above all the horizontal semantic dynamics.
- the third, finally, is a top-down reinterpretation, in which the emergent meaning of the whole impacts back on its own constituents.

These three dynamics play together and constitute an on-going unfolding of meaning that, for principle reasons, is never finished. It may, however, converge asymptotically, in which case we can get to a rather clear – although never fully well-defined picture. It may also diverge, in which case we cannot come to grips with the issue, even if we draw on constellatory logic.

In experiencing art, I would argue, this constellatory logic inevitable plays a major role. The way in which a poem "unfolds" its meaning for us can serve as a good example for the three dynamics mentioned above:

- A poem is a constellation of words. A first effect of this is that all the words shed mutually light on each other, i.e. interpret and reinterpret each other mutually. (For this reason there aren't two poems in which exactly the same "moon" would shine.) It is characteristic for a poem that the words that constitute it continue to unfold their meaning in and via their specific constellation.
- A second effect is that – out of this "horizontal" semantic dynamics that occurs between the individual notions that constitute the poem – an overall meaning emerges "vertically". This emergent over-all meaning can never be comprehensively defined, due to the ongoing "horizontal" dynamics from which it results. The emergent meaning can in some cases remain extremely ambiguous or

opaque. But, exactly this ambiguity or opaqueness is, in this case, the overarching impression that emerges.

- The third dynamics, finally, concerns the "feed-back" of this emergent overarching meaning upon its own constituents. Contrary to the prior, this is not a "bottom-up", but a kind of "top-down" dynamics – in which the overarching meaning that emerged now impacts back on its own constituent components, the meaning of the individual words and the dynamics that occur between them.

The unfolding of the meaning of a poem is in this way a highly selfreferential, and thus autogenetic, process. What has been called "logic of constellations" is the trial to understand how this unfolding actually takes place – in a paratactic predication space and drawing on these three underlying dynamics.

The next question is whether there is any phenomenological evidence from cognitive psychology or cognitive neuroscience that would correspond to this philosophically derived claim of a second, constellatory mode of thinking.

Under the heading of "dual-process accounts", but also under some other headings there seems to exist an interesting debate that points in a somewhat similar direction (see Literature). In this debate the point of departure are empirical observations of how the human brain actually executes thinking processes. The one main position in this debate is that there should be two basic modes of thinking: An elder, preconceptual mode, referred to as "system 1", and a much more recent, language-based, explicitly rational, and capacity-wise rather limited mode referred to as "system 2". In terms of the basic features attributed to the two systems these findings seem to fit quite nicely with the two cognitive approaches to reality postulated here.

The arguments developed here may also explain why the elder, in my words "constellatory" mode was not replaced in toto, but why it was just complemented by the new, ratiomorphic mode. The more recent approach is a very efficient adaptation to the factual aspect of reality – but it is, for that very reason, *structurally* incapable to deal with the statu-nascendi aspect of reality. In order to cope with this, and especially the phenomena of strong selfreferentiality and autogenesis, our thinking has to draw still on the elder, constellatory mode. This constellatory mode can, however, not only be applied to pre-conceptual mental content. Once explicit concepts are available we may look also of constellations of concepts, i.e. process them in this constellatory mode. I assume that really creative and innovative thinking is characterized by the ability to switch effortlessly, seamlessly and frequently between these two modes of thinking.

But even without this specific skill to process entire constellation of concepts in a constellatory mode, the two modes of human thinking are irreducibly interwoven, even in basic thinking operations. Given the most fundamental architecture of neurons and networks of neurons the constellatory mode seems to be much closer to the underlying, neurobiological "hardware". Due to their very structure, neurons "throw together" many heterogeneous inputs and "reduce" this richness into a much simpler output, their firing frequency. Neural networks seem optimally predisposed for "associative learning", which turns out to be a form of constellatory processing.

Ratiomorphic reasoning, instead, constitutes a rather remote possibility to use biological neural networks. Probably several "windows of opportunity" had incidentally

to be open at the same time in order to allow for its initial development.[1] Otherwise many more species should have profited massively from its spontaneous development. Compared to the underlying neurobiological machinery it seems almost a bit like dressage in horse riding: horses can walk diagonally backward, but it comes by no means naturally to them.

Listing the features of each of these two modes of thinking we immediately see their respective strength and weaknesses:

	Ratiomorphic	Constellatory
content	well-defined (at least asymptotically)	meaning unfolds
formalization	possible	impossible
place-holding (formalization)	possible	impossible
proves	mandatory, coercive	up to acceptance
truth criterion	formal correctness	authentic experience
predication	either / or	paratactic
observation	detached / external	integral part / from within
conclusions	possible	impossible
(authentic) experience	not necessary	constitutive
embodiment	optional	irreducible
aspect of time	linear-sequential	expanded present
logic	Boolean	nonboolean with dynamics of constellatory unfolding

But arguing for this fundamental complementarity of the two thinking modes, a additional question arises. Why has the ratiomorphic mode attracted almost all attention in our thinking about thought – just think about the use of the notion of logic ever since Aristotle – and the constellatory to little?

The answer could be that the ratiomorphic mode is inherently affine to the search for precise explanations drawing on well-defined distinctions and concepts, and hence the dramatic overrepresentation. Only once we learned that reality itself is not comprehensively well-defined (with this proposition I am, in a way, just paraphrasing Heisenberg's "uncertainty principle") we were forced to wonder what and how to think about reality itself. It then took actually several decades until it dawned to us (a) to which extent also all theories of modern physics are grounded on categorial foundations, (b) that these categories are not isolated entities but that they form closely interrelated apparatus, and (c) that these underlying categorial apparatus have to become an integral part of our physical theories in order to understand the specifics of the quantum physical take on reality, and in order to come – derivatively – to grips with its relation to general relativity.

[1] Once developed, presumably several strong positive selection mechanisms kicked in and this explains the – in evolutionary terms – extremely rapid development of advanced cognitive skills based on syntactic language and conceptual hierarchies. As this allowed also for the "outsourcing" of cognitive evolution into modular cultural artifacts, the whole process started to accelerate itself even further – and we got, so to speak, from throwing bones to throwing bombs in just a blink.

Heisenberg's claim that the advent of quantum physics was the most significant event in twentieth century philosophy is probably correct – one just might state it somewhat more modestly and adequately as "gave rise to". Quantum physics itself found ingenuous ways how to *handle* the statu-nascendi aspect of reality mathematically – but, it did not really *understand* what it was doing.

The statu-nascendi aspect of reality is, as we saw, inseparably linked to the principle of uncertainty respectively the phenomenon of objective indeterminacy and genuine novelty. They all imply that in a phase-space portrait of the "state of affairs" exante there remains a certain, irreducible *volume* out of which several different trajectories can emerge. Only time can tell which of these options actually materializes – which implies both, strong temporality and an "objective role" for the present in physics. The incompressibility of this volume is the footprint of objective indeterminacy.

This means also that our ex-ante ignorance is not a matter of lacking knowledge; the situation is "objectively" undecided, i.e. there is nothing that could be known. In order to cope with such a setting of "objective indeterminacy" constellatory logic comes very handy. The best we can do in such a situation is to look at the entire ensemble of components (in this case options), and their specific constellation. Given objective indeterminacy, this constellatory approach is much more appropriate than arbitrarily picking out one possibility and treating it with Boolean rigidity.[2]

All we can build-on, as long as reality is still "underdefined", i.e. still in statu-nascendi, are *constellations of options, respectively components or features.* By looking at them in the specific configuration they can "shed light on each other", i.e. mutually interpret their meaning. Confronted with such an underdefined, "objectively uncertain" situation, our cognition tries to make sense of the specific constellation of components. Traditionally this way of thinking (i.e. of linking mental content) has often been called "intuition", or "intuitive thinking" or "(gut) feeling". In not being easy to verbalize it is closely related also to (implicit) emotional and esthetic assessments.

The challenge for a logic of constellations is to address and represent, respectively approximate and mimic this unfolding. In trying this two characteristics need to be avoided: well-definedness (as it would curb the dynamics) and "everything goes" (as this would be equivalent to the "ex falso-quodlibet" catastrophe, and, in addition, it would make the notion of rules or principles superfluous in the first place).

The logic of constellations that I am proposing tries to avoid both traps by introducing the three mentioned principles of "semantic unfolding". But it should be stressed again, that this does not give us back the rigid truth criteria of classical logic. Instead, both, the authentic presence of the semantic content in the respective constellation is needed (i.e. no place-holding is possible which is inherent in all

[2] In treating this objective indeterminacy mathematically one has the huge advantage of putting all that can be said in a seemingly well-defined formalism, the development of the probability function, which is fully deterministic and time-reversible. Yet, one keeps the "real" ontological meaning absolutely open – by cramming all the uncertainty in the unsuspicious use of complex numbers. This mathematical 'trick' is very elegant and powerful, but – by offering a somewhat misleading "quasi-classicality" – it also contributed to hiding the radical break of quantum physics with the classical / factual notions of time and reality.

efforts to formalize) and the authentic experience of the unfolding of meaning are required (that is the reason why "summarizing" a poem does not work.) Authentic experience is the only and ultimate criterion of truth in a logic of constellations.

In the western occidental cultural tradition a split occurred which separated "science" – as the realm of well-defined accounts – from "art" – as the realm for which the autogenetic unfolding of meaning out of constellations is characteristic. Seen from the conceptual framework developed here, too radical a rift between these two is rather unfortunate. It hides the phenomenon that reality itself has both aspects to it, and thus requires both modes of cognition to be used in a complementary, not in a separated and dichotomized way. Only by using the two modes of thinking in their complementary way we can regain access to something that got more and more marginalized during the development of modern science and technology: the co-perception of what I would call "the objective wonderfulness of reality."

In this sense, constellatory thinking operations may still today play the dominant role in three domains of advanced cognition: (1) creativity and intuition, the latter also as an accompanying factor of profound expertise, (2) the whole realm of arts, and (3) for the experience of meaning and sense in our lives. Obviously, constellatory reasoning can be and usually is, in all three cases, massively and widely interwoven with ratiomorphic components of thinking. But that does not deny its predominance in these domains of advanced human cognition.

An account of reality that – in a self-immunizing way – focuses more and more on the well-defined aspects of reality amounts at the end to a mental and cultural situation that could be described as a "facticity imprisonment". For Midas everything he touched turned into gold, according to his own wish. At the end, this led to his death by starvation. Today we are in a comparable situation: in order to accept something as real we require it to be factual – and thus we become more and more deprived of the coperception of the genuine openness, and, thus "wonderfulness" of reality. One could even argue that the contemporary predominance of the paradigm of "power, possession and control", is the futile effort to compensate for this strongly felt, but hardly understood cognitive deficit.

Summing up, the punch line of the argument is that human thinking may be characterized by two complementary modes of linking mental content:

- In phylogenetic terms, a rather late, ratiomorphic reasoning which is based on (asymptotically) well-defined or well-definable operations that are also relatively easy to verbalize. The logical core of this mode of reasoning is Boolean logic (which turns out to be also the indispensable meta-logic of all so-called many-value, modal or temporal logics).
- The phylogenetically elder, but through the course of cognitive evolution also increasingly sophisticated mode of "constellatory logic". Being initially clearly preconceptual, this mode of linking mental content is reutilized again on the level of concept-based thinking. Good examples for this are esthetic or "intuitive" judgments. Even if applied on the level of conceptual thinking, the constellatory

operations are still constitutively not well-defined and are, therefore, not formalizable. This is the inevitable price that has to be paid for being able to address strongly selfreferential and autogenetic phenomena, i.e. the statu-nascendi aspect of reality.[1] In the history of philosophy there are two – rather controversial – thinkers whose thinking was essentially based on a much richer understanding of "logic", Heraclites and Hegel. One can, however, read almost the entire history of philosophy under the aspect how a complementarity of thinking modes is alluded to, at least implicitly.

References

Evans JStBT (2003) In two minds – dual-processing accounts of reasoning. TREND Cogn Sci 7:10

Evans JStBT (2008) Dual-processing accounts of reasoning, judgement and social cognition. Ann Rev Psychol 59:255–278

Regarding the claim of two complementary categorial apparatus and their correspondence to the factual and the statu-nascendi aspect of reality see: von Müller AAC (1983) Zeit und Logik. Wissenschaftszentrum München (ed) Wolfgang Bauer Verlag, München

Filk T, von Müller A (2009) Quantum physics and consciousness: the quest for a common conceptual foundation. Mind Matter 7(1):59–80

[1] A final remark on further literature: It would have been beyond the scope of this paper - and, even more so, beyond my limited competences - to systematically introduce and discuss the many approaches to "complementary modes of thinking" that exist in the history of philosophy and, more recently, in cognitive science. Even a not too detailed overview would probably require a book of its own. From the history of philosophy I would just like to mention three – rather controversial - thinkers whose thinking was essentially based on a much richer understanding of the notion of "logic", Heraclites, Hegel and Heidegger. One can, however, read almost the entire history of philosophy under the aspect how a complementarity of thinking modes was alluded to, explicitly or implicitly. Regarding modern "dual-processing" accounts see for an interesting overview Evans, J.St.B.T. (2003): In two minds – dual-processing accounts of reasoning. TRENDS in Cognitive Sciences Vol. 7 No. 10 and more recently Evans, J.St.B.T. (2008): Dual-Processing Accounts of Reasoning, Judgement and Social Cognition. Annual Review of Psychology 59, 255-278. Very interesting and relevant contributions to the issue have been provided also by Vinod Goel (see: Goel, V. (1995): Sketches of thought. Cambridge, MA: MIT Press; Vartanian, O. & Goel, V. (2004): Neuroanatomical correlates of aesthetic preference for paintings. NeuroReport, Vol. 15, No. 5, pp. 893-897; Vartanian, O. & Goel, V. (2005): Neural Correlates of Creative Cognition. In: C. Martindale, P. Locher, & V. Petrov (Eds.), Evolutionary and neuro-cognitive approaches to the arts. Baywood Publishing; Goel, V., Buchel, C., Frith, C., Dolan, R. (2000): Dissociation of Mechanisms Underlying Syllogistic Reasoning. NeuroImage, Vol. 12, No. 5, pp. 504-514). Regarding the claim of two complementary categorial apparatus and their correspondence to the factual and the statu-nascendi aspect of reality see: von Müller, A.A.C (1983): Zeit und Logik. Wissenschaftszentrum München (ed.), Wolfgang Bauer Verlag, Munich and more recently Filk, T. & von Müller, A. (2009): Quantum Physics and Consciousness: The Quest for a Common Conceptual Foundation. Mind and Matter 7/1, 59-80.

Part II
Components of Thinking

Categorization and Object Shape

Markus Graf

Abstract Categorization is essential for perception and provides an important foundation for higher cognitive functions. In this review, I focus on perceptual aspects of categorization, especially related to object shape. In order to visually categorize an object, the visual system has to solve two basic problems. The first one is how to recognize objects after spatial transformations like rotations and size-scalings. The second problem is how to categorize objects with different shapes as members of the same category. I review the literature related to these two problems against the background of the hierarchy of transformation groups specified in Felix Klein's *Erlanger Programm*. The *Erlanger Programm* provides a general framework for the understanding of object shape, and may allow integrating object recognition and categorization literatures.

1 Introduction

Categorization is regarded as one of the most important abilities of our cognitive system, and is a basis for thinking and higher cognitive functions. An organism without such abilities would be continually confronted with an ever-changing array of seemingly meaningless and unrelated impressions. The categorization of environmental experiences is a basic process that must be in place before any organism can engage in other intellectual endeavors. In the words of the cognitive linguist George Lakoff: "There is nothing more basic than categorization to our thought, perception, action, and speech. Every time we see something as a kind of thing, for

M. Graf
Department of Psychology, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a, 04103, Leipzig, Germany and Max Planck Institute for Biological Cybernetics, Tübingen, Germany
e-mail: markus.graf@cbs.mpg.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, *On Thinking*,
DOI 10.1007/978-3-642-03129-8_6,

example, a tree, we are categorizing. Whenever we reason about kinds of things – chairs, nations, illnesses, emotions, any kind of thing at all – we are employing categories. (…) Without the ability to categorize, we could not function at all, either in the physical world or in our social and intellectual lives. An understanding of how we categorize is central to any understanding of how we think and how we function …" (Lakoff 1987, pp. 5–6).

A fundamental aspect of categorization is to visually recognize and categorize objects. Probably the most important feature for object categorization is object shape (e.g., Biederman and Ju 1988). Usually we are able to visually recognize objects although we see them from different points of view, in different sizes (due to changes in distance), and in different positions in the environment. Even young children recognize objects so immediately and effortlessly that it seems to be a rather ordinary and simple task. However, changes in the spatial relation between the observer and the object lead to immense changes of the image that is projected onto the retina. Hence, to recognize objects regardless of orientation, size, and position is not a trivial problem, and no computational system proposed so far can successfully recognize objects over a wide range of categories and contexts. The question about how we recognize objects despite spatial transformations is usually referred to as the first basic problem of object recognition. Moreover, we are not only able to recognize identical objects, after spatial transformations, but we can also effortlessly categorize an unfamiliar object, for instance a dog, a bird, or a butterfly, despite, sometimes, large shape variations within basic categories (e.g., Rosch et al. 1976). How do we generalize over different instances of an object class? This ability for class-level recognition or categorization is considered as the second basic problem of recognition. The main difficulty in classification arises from the variability in shape within natural classes of objects (e.g., Ullman 2007).

Objects can be recognized or categorized on different levels. For example, a specific object can be categorized as an *animal*, as a *dog*, as a *beagle*, or as my dog Snoopy. One of these levels has perceptual priority, and is called the basic level of categorization (Rosch et al. 1976; for a review see e.g., Murphy 2002). The basic level is usually also the entry level of categorization (Jolicoeur et al. 1984). Thus, we tend to recognize or name objects at the basic level, i.e., we see or name something as *dog*, *cat*, *car*, *table*, *chair*, etc. The level above the basic level is called superordinate level (e.g., vehicle, animal), while the level below the basic level is called subordinate level (limousine, van, hatchback or collie, dachshund, beagle, etc.).

The basic level is the most inclusive level at which members of this category have a high degree of visual similarity, and at which observers can still recognize an average shape, created from the shapes of several category members. (Rosch et al. 1976).[1] Therefore the basic level is the highest level of abstraction at which it is possible to form a mental image which is isomorphic to an average member of the class

[1] In addition, the basic level is the most inclusive level at which we tend to interact with objects in a similar way (Rosch et al. 1976), indicating the importance of knowledge about motor interactions for categorization (see Helbig et al. 2006).

and, thus, the most abstract level at which it is possible to have a relatively concrete image of a category (Rosch et al. 1976, Exp. 3 and 4; see also Diamond and Carey 1986). The pictorial nature of category representations up to the basic level has been confirmed in further experiments. In a signal detection experiment, subjects were better in detecting a masked object when the basic level name of the object was given prior to each trial. Superordinate level names did not aid in detection, which suggests that superordinate level categories are not represented in a pictorial code. Moreover, a priming experiment demonstrated that the basic level is the most abstract level at which preceding exposure of the category name affected same responses under physical identity instructions (Rosch et al. 1976, Exp. 5 and 6). These findings suggest that the basic level is the highest level at which category representations are image-based or pictorial.

Evidence for image-based or pictorial representations has been found also in the object recognition literature. The majority of findings indicate that recognition performance depends systematically on the amount of transformation (rotation, size-scaling, and shift in position) to align input and memory representations (for review see Graf 2006; for details see Sect. 3). This dependency suggests that representations are in a similar format as the visual input, whereas abstract representations should – by definition – be independent of image transformations.

A number of different accounts of recognition and categorization have been proposed, differing in the abstractness of the postulated representations (for reviews see Edelman 1997, 1999; Graf 2006; Murphy 2002; Palmeri and Gauthier 2004; Ullman 1996). Several models rely on relatively abstract representations. Models from the categorization literature are usually based on abstract features or properties (e.g., Nosofsky 1986; Cohen and Nosofsky 2000; Maddox and Ashby 1996; Markman 2001), but abstract features seem to be limited in their capacity to describe complex shapes. Structural description models from the recognition literature involve a decomposition into elementary parts and categorical spatial relations between these parts (like above, below, side-of), and thus are based on abstract propositional representations (e.g., Biederman 1987; Hummel and Biederman 1992; Hummel and Stankiewicz 1998). However, models that rely on abstract representations are difficult to reconcile with strong evidence for a systematic dependency on image transformations, like rotations and size-scalings (for review see, e.g., Graf 2006).

Several different image-based approaches have been proposed, which are better suited to account for the dependency on transformations (for reviews see Jolicoeur and Humphrey 1998; Tarr 2003). Early alignment models relied on transformational compensation processes, like mental rotation, in order to align stimulus representations and memory representations (e.g., Jolicoeur 1985, 1990a; Ullman 1989, 1996). As evidence has accumulated against mental rotations in object recognition (Jolicoeur et al. 1998; Willems and Wagemans 2001; Farah and Hammond 1988; Gauthier et al. 2002; for a review see Graf 2006), later image based approaches avoided the notion of transformation processes (e.g., Edelman 1997, 1998; Edelman and Intrator 2000, 2001; Perrett et al. 1998; Riesenhuber and Poggio 1999; Ullman 2007).

Moreover, hybrid models have been proposed in an attempt to combine structural and image-based approaches (Foster and Gilson 2002; Hayward 2003). Some hybrid

models have been derived from structural description models (Hummel and Stankiewicz 1998; Thoma et al. 2004), while others are extensions of image-based models (Edelman and Intrator 2000, 2001). The notion of structured representations is implicit also in the structural alignment approach brought forward in the literatures on similarity, analogy, and categorization (e.g., Medin et al. 1993; Gentner and Markman 1994, 1995; Goldstone and Medin 1994; Goldstone 1994a, 1994b, 1996; Markman 2001). This approach combines structured representations with the notion of alignment, the latter being used also in image-based approaches of recognition (Ullman 1989, 1996; Lowe 1985, 1987). More recently, a transformational framework of recognition involving alignment has been proposed, now relying not on mental rotations, but on an alignment based on coordinate transformations (Graf 2006; Graf et al. 2005; Salinas and Sejnowski 2001; Salinas and Abbott 2001; for further transformational approaches see Hahn et al. 2003; Leech et al. 2009).

Although the terms *recognition* and *categorization* are often used synonymously, they are investigated in two separate research communities. Traditionally the term *recognition* is more associated with perception and high-level vision, while *categorization* is more associated with cognition (e.g., Palmeri and Gauthier 2004). Members of both communities attend to different conferences, with relatively little overlap (see Farah 2000, p. 252). Surprisingly, relatively little research in the field of categorization is related to object shape (for reviews see Murphy 2002; Ashby and Maddox 2005). Consequently, relatively few attempts were made in order to come to an integrative approach of recognition and categorization (for exceptions see Edelman 1998, 1999; Nosofsky 1986; for an integrative review see Palmeri and Gauthier 2004). The aims of this article are related to this shortcoming. First, I will propose that Felix Klein's hierarchy of transformation groups can be regarded as a framework to conceptualize object shape and shape variability within categories up to the basic level. Second, given the commonalities between object recognition and categorization, I will lay out the foundations for an integrative transformational framework of recognition and categorization.

2 Form and Space

A prevalent idea in present cognitive neuroscience is that shape information and spatial information are processed in different visual streams, and therefore more or less dissociated. While shape processing for object recognition is postulated to occur exclusively in the ventral stream, the dorsal visual stream is involved in spatial tasks (Ungerleider and Mishkin 1982; Ungerleider and Haxby 1994), or perception for action (Milner and Goodale 1995; Goodale and Milner 2004). However, from a logical or geometrical point of view, shape and space cannot be strictly separated, but are closely related. As Stephen Kosslyn (1994, p. 277) argued, a shape is equivalent to a pattern formed by placing points (or pixels) at specific locations in space; a close look at any television screen is sufficient to convince anyone of this observation. Thus, shape is nothing more than a set of locations occupied by an object (Farah 2000, p. 71). Moreover, observers need to recognize shapes independent of the spatial

relation between observer and object, and thus compensate or account for spatial transformations in object recognition. Given this tight connection between form and space, it seems reasonable to investigate whether a geometrical (spatial) theory of shape is feasible. There is a growing body of evidence suggesting that areas in the parietal cortex – that is, areas usually associated with spatial or visuomotor processing – are involved in the recognition of disoriented objects (Eacott and Gaffan 1991; Faillenot et al. 1997, 1999; Kosslyn et al. 1994; Sugio et al. 1999; Vuilleumier et al. 2002; Warrington and Taylor 1973, 1978). A recent experiment using transcranial magnetic stimulation confirmed that the parietal cortex is involved in object recognition (Harris et al. 2009). Interestingly, the categorization of distorted dot pattern prototypes (in which different dot patterns were created by shifting the dots in space) involves not only typical shape-related areas like lateral occipital cortex, but also parietal areas (Seger et al. 2000; Vogels et al. 2002).

In accordance with the close connection between shape and space, I will argue that shape and shape variability can be conceptualized in terms of geometrical transformations. The organization of these different types of transformations can be described by *Felix Klein's (*1872*/1893) Erlanger Programm*, in which Klein proposed a nested hierarchy of geometrical transformation groups to provide an integrative framework for different geometries. This hierarchy of transformation groups ranges from simple transformations like rotations, translations (shifts in position), reflections, and dilations (size-scalings) – which make up the so-called Euclidean similarity group – to higher (and more embracing) transformation groups, namely affine, projective and topological transformations. I will explain and illustrate these transformations below, but before I will provide a brief historical survey to shed some light on the importance of the *Erlanger Programm* for geometry.

In the nineteenth century, geometry was in danger of falling apart into several separate areas, because different non-Euclidean geometries have been developed by mathematicians like Gauß, Lobatschewsky, and Bolayi, and later elaborated by Riemann. In 1872, Felix Klein was appointed as an ordinary professor of mathematics in Erlangen. In his inaugural address he proposed that different geometries can be integrated into one general framework – a project which was later called the *Erlanger Programm*. Klein argued that geometrical properties and objects are not absolute, but are relative to transformation groups.[2] For instance, a circle and an ellipse are different objects in Euclidean geometry, but in projective geometry all conic sections are equivalent. Regarding the projective group, a circle can be easily transformed into an ellipse or any other conic section.

Based on the idea that a geometry is defined relative to a transformation group, it was possible to integrate the different geometries by postulating a nested hierarchy of transformation groups. Euclidean geometry, projective geometry and space-curving geometries simply refer to different geometrical transformation groups in Klein's

[2] In mathematics, a group is a set that has rules for combining any pair of elements in the set, and that obeys four properties: closure, associativity, existence of an identity element and an inverse element. The mathematical concept of group has been used in cognitive psychology (e.g., Bedford 2001; Chen 2005; Dodwell 1983; Leyton 1992; Palmer 1983, 1989; Shepard 1994).

hierarchy. Note that these transformation groups are also important for the understanding of object shape.

The first important group in the hierarchy of transformations here is the so-called *Euclidean similarity group*, which is made up of rotations, translations (shifts in position), reflections, and size-scalings. The Euclidean similarity group can be regarded as the basis of Euclidean geometry (see Ihmig 1997).[3] Mainly these transformations need to be compensated after changes in the spatial relation between observer and object.

The next higher transformation group is the group of *affine* transformations, which includes transformations like linear stretchings or compressions in one dimension, and also linear shear transformations, which change the angle of the coordinate system. In short, affine transformations are linear transformations that conserve parallelism, i.e., in which parallel lines remain parallel. A simple affine stretching transformation occurs when TV programs in the usual 4:3 format are viewed on the new 16:9 TV sets. Affine transformations are nicely illustrated by Albrecht Dürer (1528)/(1996), who was probably the first who systematically investigated the influence of geometrical transformations on object shape, focusing on human bodies and faces (see Fig. 1). As every higher transformation group

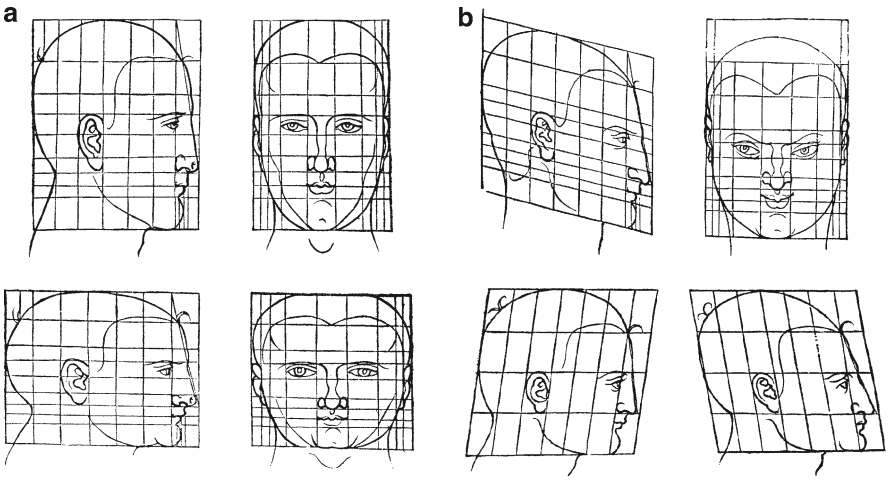

Fig. 1 As demonstrated by Albrecht Dürer (1528/1996), affine transformations (i.e., linear transformations that conserve parallelism) provide a way to account for some of the shape differences between different heads. (**a**) *On the left side* affine compression and stretching transformations are depicted. (**b**) *On the right* affine shear transformations are shown, which can include a transformation of the angle of the coordinate system (while parallels still remain parallel). *Note*: Drawings by Albrecht Dürer 1528, State Library Bamberg, Germany, signature L.art.f.8a. Copyright by State Library Bamberg, Germany. Adapted with permission

[3] The Euclidean similarity group can be further subdivided (e.g., Bedford 2001), but this is not important for present purposes. A more detailed description of the hierarchy of transformation groups can be found in Michaels and Carello (1981, p. 30–37), Palmer (1983), or Cutting (1986).

includes the lower group (nested hierarchy), the Euclidean similarity group is included in the group of affine transformations.

The next group in the hierarchy is the group of *projective* transformations. Projective transformations are linear transformations which do not necessarily conserve the parallelism of lines (and therefore violate Euclid's parallelity axiom). Central perspective in Renaissance paintings is constructed on the basis of projective geometry, as parallel lines intersect (in the vanishing point). Moreover, projective transformations allow describing yet further systematic changes of object shape beyond affine transformations (see Fig. 2).

Finally, the highest (or most basic) group in the hierarchy of point transformations is the group of *topological* transformations.[4] Topological transformations allow for nonlinear transformations, so that straight lines can be transformed into curved lines. Topological transformations can be illustrated by deforming a rubber sheet without ripping it apart. For this reason, topological geometry is often called

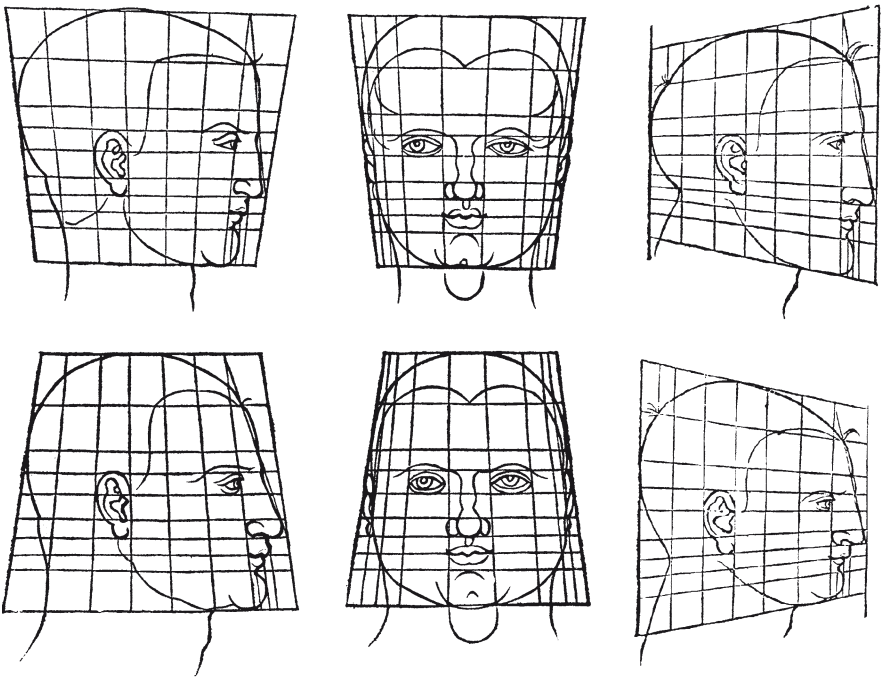

Fig. 2 Projective transformations of shapes, which allow for linear transformations that violate parallelism, can account for a still larger range of shape variations between different heads. *Note*: Drawings by Albrecht Dürer 1528, State Library Bamberg, Germany, signature L.art.f.8a. Copyright by State Library Bamberg, Germany. Adapted with permission

[4]There are also transformations which go beyond point transformations (see Ihmig 1997). However, these do not seem to be of primary importance for an understanding of object shape.

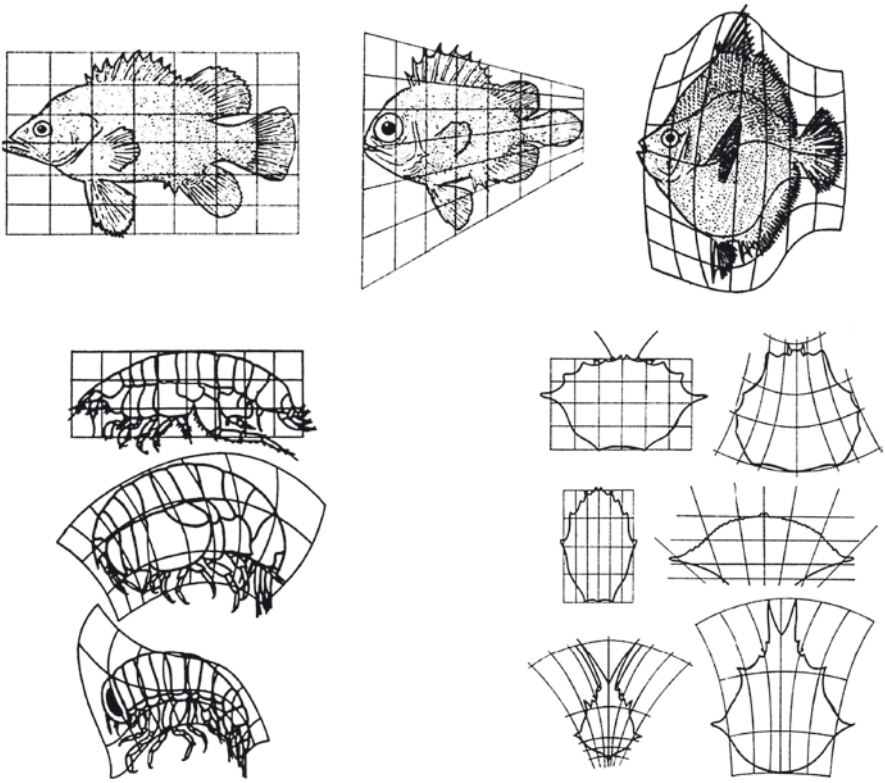

Fig. 3 Topological transformations, which include also nonlinear (deforming) transformations, describe variations of the shapes of closely related animals, like different types of fish. *Note*: Drawings by Thompson 1917. Copyright by Cambridge University Press. Adapted with permission

rubber sheet geometry. As the hierarchy of transformation groups is a nested hierarchy, the group of topological transformations includes not only space-curving transformations, but all transformations that were described above. Topological geometry was first employed within a scientific theory in 1915 in Einstein's General theory of relativity. Only 2 years later, Thompson (1917)/(1942) used topological transformations to account for differences in the shapes of closely related animals (Fig. 3).

These different geometrical transformations provide a systematic way to describe the shape variability of biological objects, like facial profiles. A specific type of topological transformation is suited to characterize the remodeling of facial profiles by growth (Pittenger and Shaw 1975; Shaw and Pittenger 1977; see Fig. 4). This transformation has been dubbed "cardioidal" transformation, because it changes a circle into a heart shape. Cardioidal transformations can be described mathematically by using rather simple trigonometric functions. A matrix of head profiles was created by applying the cardioidal and an affine shear transformation

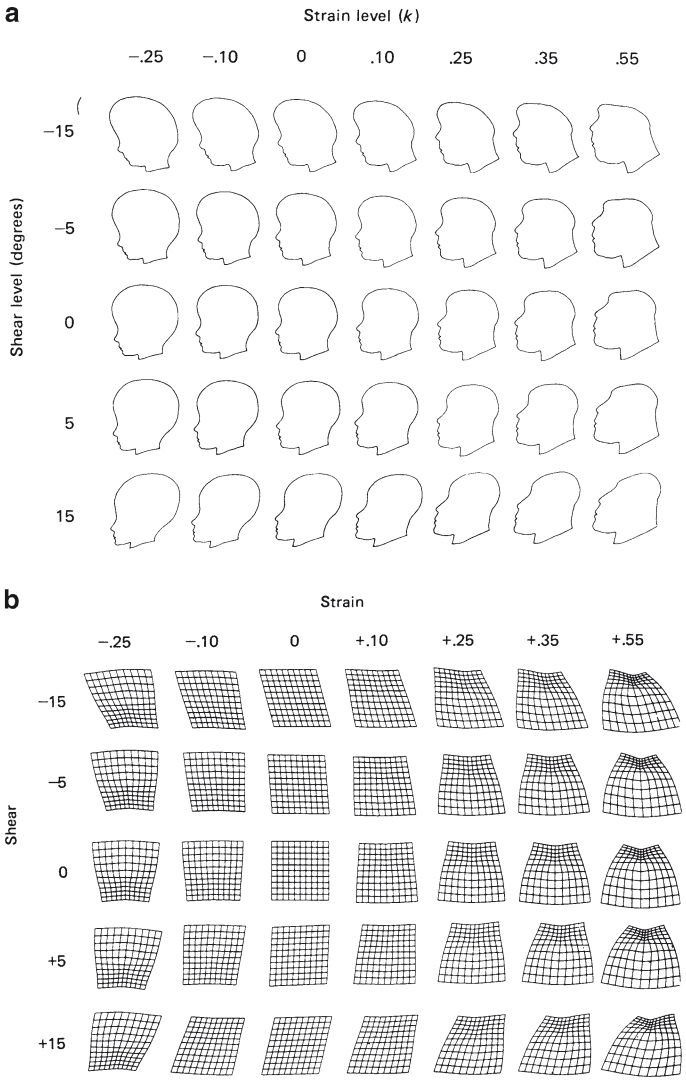

Fig. 4 Affine shear and topological transformations provide a principled description of the shape space of the category *head*. (**a**) A set of facial profiles was created with topological transformations (*within rows*) and affine shear transformations (*within columns*), in order to investigate age perception (Shaw and Pittenger 1977). (**b**) The geometrical transformations can be conceptualized as transformations of the underlying coordinate system, i.e., as transformations of space. The standard grid is at shear = 0, strain = 0. Note that the deformations underlying the shape changes are rather simple. *Note*: Figures by Shaw and Pittenger 1977. Copyright by Robert Shaw. Adapted with permission

to the profile of a 10-year-old boy. The cardioidal (nonlinear) transformation allows deforming the head into the profile of a baby, into the profile of a grown-up man, or into some Neanderthal-man-like profile (see Fig. 4a, horizontal rows). Within

each column, the level of affine shear is modified. The underlying transformations that correspond to these shape changes are visualized by the deformation of the corresponding coordinate systems (Fig. 4b).[5] Thus, the shape changes are created by a deformation of the underlying space.

More recently, these ideas have been extended by proposing that the shape variability of members of a given basic-level category can be described by topological transformations (Graf 2002). The different facial profiles in Fig. 4a can be considered not only as phases in a growth process, but also as different members of the category *head*. By allowing for topological transformations, shape variability within categories can be accounted for – up to the basic level of categorization (Graf 2002). This proposal is consistent with the finding that the basic level is the highest level at which members of a category have similar shapes (Rosch et al. 1976), and it provides a systematic way to deal with these shape differences. This approach works well for biological objects, and also for many artifact categories (for constraints see Graf 2002). Topological transformations correspond to morphing in computer graphics. With morphing, the shape of one object is transformed into another, based on an alignment of corresponding points or parts. The use of topological (morphing) transformations has significant advantages: First, morphing offers the possibility to create highly realistic exemplars of familiar categories, moving beyond the artificial stimuli previously used in visual categorization tasks (for review see Ashby and Maddox 2005). Second, morphing is an image transformation which aligns corresponding features or parts. Thus, morphing is both image-based and structural. Third, the method allows to vary the shape of familiar objects in a parametric way, and thus permits a systematic investigation of shape processing.

The framework of the *Erlanger Programm* has been employed before in theories in perceptual and cognitive psychology. The *Erlanger Programm* and mathematical group theory proved to be useful to account for perceptual constancy (e.g., Wagemans et al. 1997; Cassirer 1944), perceptual organization (Chen 2001, 2005; Palmer 1983, 1989, 1999), the perception of motion and apparent motion (e.g., Chen 1985; Shepard 1994; Palmer 1983; Foster 1973, 1978), the perception of age (e.g., Shaw and Pittenger 1977), event perception (Warren and Shaw 1985), object identity decisions (Bedford 2001), and object categorization (Graf 2002; Shepard 1994). The *Erlanger Programm* may provide a basis for a broad framework of visual perception, which covers not only recognition and categorization, but also perceptual organization. Thus, the *Erlanger Programm* seems to be a promising theoretical framework, considering that an integral theory of perceptual organization and categorization is necessary (Schyns 1997).

In previous approaches the transformation groups of the Erlanger Programm have been typically used as a basis to define invariants, i.e., formless mathematical properties which remain unchanged despite spatial transformations (e.g., Gibson 1950; Todd et al. 1998; Van Gool et al. 1994; Wagemans et al. 1996; see already

[5]Nonlinear transformations seem to play a role also within the visual system. The retina is not flat but curved, and projections onto the retina are therefore distorted in a nonlinear way. Moreover, the projection from the retina to the primary visual cortex is highly nonlinear.

Cassirer 1944; Pitts and McCulloch 1947; for a review see Ullman 1996). For instance, the cross ratio is a frequently used invariant of the projective group (e.g., Cutting 1986; a description of the cross ratio can be found in Michaels and Carello 1981, pp. 35–36).[6] Invariant property approaches may be mathematically appealing but have at least two severe problems: The higher the relevant transformation group is in Klein's hierarchy, the more difficult it gets to find mathematical invariants (e.g., Palmer 1983). And, more important, the invariants that were postulated to underlie object constancy in the visual system could often not be empirically confirmed (e.g., Niall and Macnamara 1990; Niall 1992; but see Chen 2005). In the next section I will review evidence demonstrating that recognition and categorization performance is not invariant, but depends on the amount of geometrical transformation.

3 Recognition and Categorization Performance Depend on Spatial Transformations

As I argued in Sect. 2, Felix Klein's hierarchy of transformation groups offers a general way to conceptualize shape and shape variability. It provides an excellent framework for reviewing the existing studies both on shape recognition and categorization. Most studies in the object recognition literature are related to transformations of the Euclidean similarity group, especially to *rotations*, *dilations* (size-scalings) and *translations* (shifts in position). These transformations are the ones most relevant for recognizing a specific object, because the *same* object may be encountered in different orientations, positions, and sizes. Although we are able to recognize objects after spatial transformations, reaction times (RTs) typically increase with increasing transformational distance. This has been shown extensively for orientation, both in the picture plane (e.g., Jolicoeur 1985, 1988, 1990a; Lawson and Jolicoeur 1998, 1999) and in depth (e.g., Lawson and Humphreys 1998; Palmer et al. 1981; Srinivas 1993; Tarr et al. 1998; Lawson et al. 2000; for reviews see e.g., Graf 2006; Tarr 2003). There is also plenty of evidence that recognition performance depends on size (e.g., Bundesen and Larsen 1975; Bundesen et al. 1981; Cave and Kosslyn 1989; Jolicoeur 1987; Larsen and Bundesen 1978; Milliken and Jolicoeur 1992; for a review see Ashbridge and Perrett 1998). Moreover, an increasing number of studies show position dependency (Dill and Edelman 2001; Dill and Fahle 1998; Foster and Kahn 1985; Nazir and O'Regan 1990; Cave et al. 1994). Neurophysiological studies show a similar dependency on orientation, size and position (for a review see Graf 2006).

[6] Note that invariants need not necessarily be defined in relation to mathematical groups; invariants can be defined also regarding perspective transformations (e.g., Pizlo 1994), which do not fulfill the requirements of a mathematical group.

What about the higher transformation groups in Klein's hierarchy? Also for specific *affine transformations*, like stretching or compressing in one dimension, a monotonic relation between the extent of transformation and performance was found. The dependency on the amount of affine transformations has been demonstrated for simple shapes like ellipses (Dixon and Just 1978). Two ellipses were presented simultaneously, varying in shape by an affine stretching or compression. Subjects were instructed to judge whether the two ellipses were identical either regarding height or width (the relevant dimension was indicated before each trial). RTs deteriorated systematically with increasing affine stretching or compression of the ellipses – even though just the irrelevant dimension has been transformed. Dixon and Just argued that the stimuli were compared via a normalization process analogous to mental rotation and size scaling. Using more realistic stimuli, William Labov (1973) presented line drawings of cup-like objects that were created by changing the ratio of width to height (an affine stretching or compression). This manipulation changed the shape of the cup, and made it more mug-like, vase-like or bowl-like. The likelihood of assigning the objects into these categories varied with context. For instance, the likelihood of categorizing a cup-like object as a vase, for example, increased in the context "flower" (as compared to the "coffee" context), indicating that category boundaries are at least to some degree vague and context-dependent. More interesting here, Labov's findings also indicate that the likelihood of assigning an object to a category changes with the ratio of width to height – which may be regarded as first evidence that categorization is influenced by affine transformations. Further evidence that categorization performance depends on affine transformations can be found in a study by Cooper and Biederman (1993), although it was not designed to investigate this issue.

Up to today, relatively little research has been done regarding *projective transformations*, i.e., linear transformations which do not necessarily conserve the parallelism of lines. Evidence for a monotonic relation between the extent of projective transformation and task performance was found in an experiment which was designed to investigate whether the visual system distinguishes between different types of projective transformations (Wagemans et al. 1997; see also Niall 2000). Wagemans et al. presented three objects simultaneously on a computer screen, one on top as a reference stimulus and two below it. Subjects were instructed to determine which of the two patterns best matched the reference pattern. Recognition accuracy deteriorated with increasing amount of projective transformation. Thus, there is some provisional evidence that recognition performance depends also on the amount of projective transformations. A dependency on affine and projective transformations has also been demonstrated in neurophysiological experiments. The neural response of shape-tuned neurons in IT depends systematically on the amount of affine and projective transformations (Kayaert et al. 2005).

Furthermore, *topological (i.e., space-curving) transformations* play a role in visual perception. For instance, topological shape transformations were investigated in age perception (e.g., Pittenger and Shaw 1975; Shaw and Pittenger 1977; Pittenger et al. 1979; Mark and Todd 1985). Subjects had to age-rank facial profiles that were subjected to different amounts of affine shear or topological (cardioidal)

transformation (see Fig. 4). The results indicated that the age-rankings increased monotonically with increasing amount of topological (cardioidal) transformation. In addition, profiles transformed by the cardioidal transformation elicited more reliable rank-order judgments than those transformed by affine shear transformation. These experiments supplied evidence that age perception involves topological transformations. These transformations work not only for 2D shapes, but similar results have been found for 3D heads (e.g., Bruce et al. 1989).

Also object categorization is systematically related to the amount of topological transformation. A series of experiments with dot patterns, that were conducted to investigate the formation of abstractions or prototypes (e.g., Posner and Keele 1968, 1970; Posner et al. 1967; Homa et al. 1973), show a systematic dependency on nonlinear deforming transformations. In these experiments random dot patterns were defined as prototypes and distorted with statistical methods in order to produce different exemplars of the same category. The results indicate a monotonic relation between the extent of distortion and dependent variables like RT and error rate. These statistical distortions can be understood as topological transformations, if one assumes that the space between the dots is distorted: Imagine that the dots are glued onto a rubber sheet which can be stretched or compressed in a locally variable way. Thus, the dot pattern experiments fit nicely within an account suggesting that deforming transformations are involved in object categorization.

Further evidence for a monotonic relation between the RT and the amount of topological transformations can be derived from experiments by Edelman (1995) and Cutzu and Edelman (1996, 1998), using animal-like novel objects. A number of shapes were created by varying shape parameters that lead to nonrigid (morphing) transformations of the objects. The shape parameters were selected so that the animal-like objects in the distal shape space corresponded to a specific configuration in proximal shape (parameter) space, e.g., a cross, a square, a star or a triangle (see Fig. 5a). These configurations in proximal shape space could be recovered by multidimensional scaling (MDS) of subject data, using RT-data from a delayed matching to sample task, or similarity ratings (Fig. 5b).[7]

Additional evidence comes from studies on facial expression. Facial movements, such as smiles and frowns, can be described as topological transformations of the face. The latencies for the recognition of the emotional facial expressions increase with increasing topological transformation of facial expression prototypes (Young et al. 1997). A monotonic relation between the amount of topological transformation and RTs was also demonstrated in a task in which two faces were morphed, and the morphs had to be classified as either person A or B: The latencies for the classification of the faces increased with increasing distance from

[7] However, Edelman (1998) and Cutzu and Edelman (1996, 1998) do not interpret these results in terms of nonrigid transformations of pictorial representations, but regard the data as evidence for the existence of a low-dimensional monotonic psychological space, in which the similarity relations of the high-dimensional distal shape-space are represented. For a discussion of Edelman's account of recognition and categorization see Graf (2002).

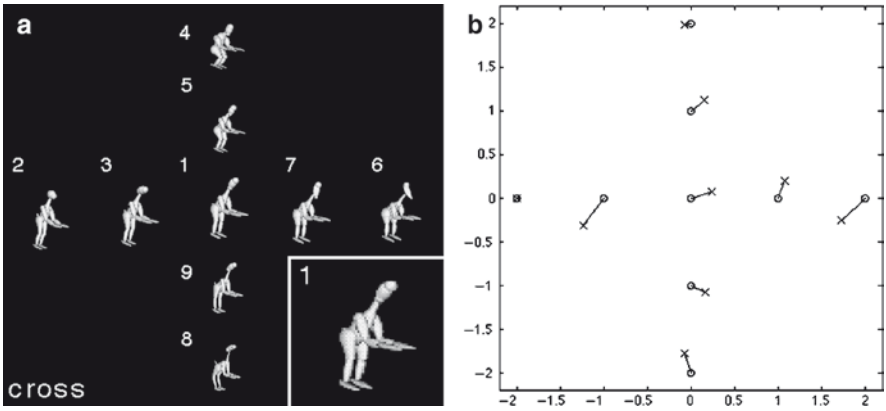

Fig. 5 A systematic dependency between recognition performance and morph transformation has been demonstrated with animal-like novel objects (Cutzu and Edelman 1996). (**a**) Subjects were confronted with several classes of computer-rendered 3D animal-like shapes, arranged in a complex pattern (here: *cross*) in a common parameter space. (**b**) Response time and error rate data were combined into a measure of perceived pairwise shape similarities, and the object to object proximity matrix was submitted to nonmetric MDS. In the resulting solution, the relative geometrical arrangement of the points corresponding to the different objects reflected the complex low-dimensional structure in parameter space that defined the relationships between the stimulus classes. *Note*: Figures by Cutzu and Edelman 1996. Copyright by Cambridge University Press. Adapted with permission

the reference exemplar (Schweinberger et al. 1999, Exp. 1a). Moreover, the study by Schweinberger et al. indicated that the increase of identification latencies is not just due to an unspecific effect of the morphing procedure, like a loss of stimulus quality, because a morphing along the emotion dimension (happy vs. angry), which was irrelevant for this task, did not affect the latencies for face identification.

More recently, topological transformations in categorization have been studied in a more direct and systematic way, using various recognition and categorization tasks (Graf 2002; Graf et al. 2008; Graf and Bülthoff 2003). Graf and collaborators systematically varied the shapes of familiar biological and artifact categories by morphing between two exemplars from the same basic category (Fig. 6). Their results indicate that categorization performance depends on the magnitude of the topological transformation between two test stimuli (Graf 2002; Graf et al. 2008). That is, when two objects were presented sequentially, performance deteriorated systematically with increasing topological (morph) distance between the two category members. These results could not be reduced to alternative accounts, such as to affine transformations, or to changes in the configuration of parts (Graf 2002; Graf et al. 2008). First, the systematic dependency was also found for categories whose exemplars differed by only small affine changes, i.e., by only small changes of the aspect ratio of the objects. Second, the dependency appeared for categories whose members have very similar part configurations. Moreover,

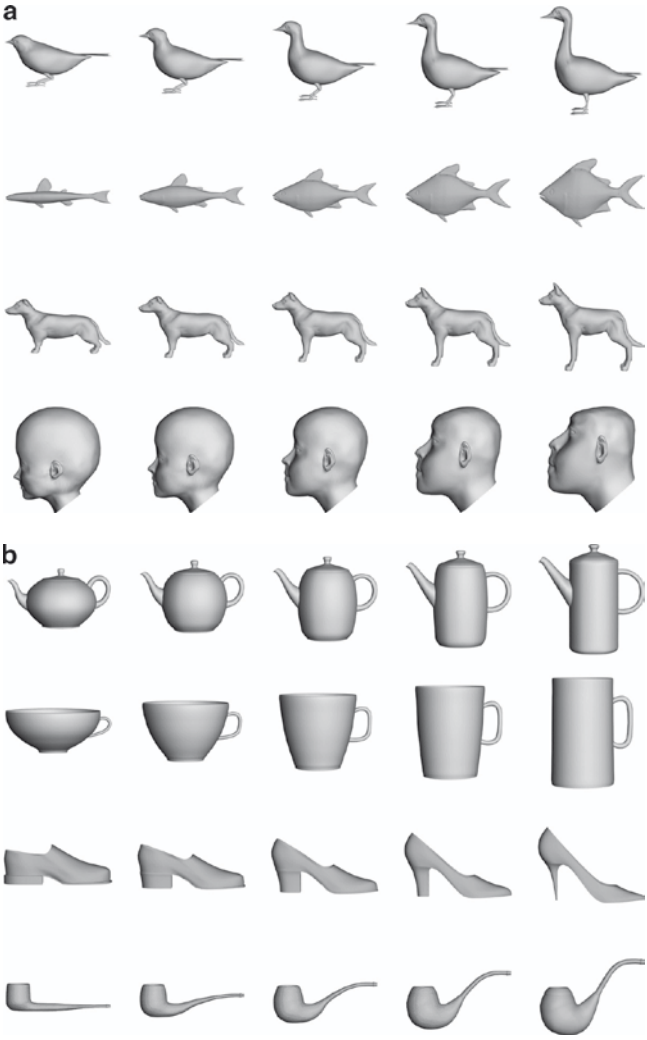

Fig. 6 Shape variability within basic-level categories can be described well with nonlinear (topological) transformations. Intermediate category members are created by morphing between two exemplars from the same basic level category. Morphing describes within-category shape variability well for biological objects and for many (but not all) artifacts (see Graf 2002, for constraints)

transformation times in the categorization task were sequentially additive, suggesting that categorization relied on analog deforming transformations, i.e., on transformations passing through intermediate points on the transformational path (Graf 2002). The effect of topological distance holds for objects that were rotated in the picture-plane, and for objects that were shifted in position (Graf and

Bülthoff 2003). A systematic dependency on the degree of topological transformation has been demonstrated also in (nonspeeded) similarity- and typicality-rating tasks (Graf 2002; Graf et al. 2008; Hahn et al. 2009). Moreover, similarity judgments were biased by the direction of a morph animation which directly preceded the similarity decision (Hahn et al. 2009). Participants were shown short animations morphing one object into another from the same basic category. They were then asked to make directional similarity judgments – "how similar is object A to object B?" for two stationary images drawn from the morph continuum. Similarity ratings for identical comparisons were higher when reference object B had appeared before object A in the preceding morph sequence. Together, these results indicate that categorization performance depends systematically both on morph distance and on the direction of transformation. These findings are consistent with deformable template matching models of categorization (Basri et al. 1998; Belongie et al. 2002).

There is also neuropsychological evidence for deforming shape transformations in the visual system. A patient with intermetamorphosis reported that animals and objects she owned took the form of another animal or object. She also experienced changes in her husband's appearance, which could become exactly like that of a neighbor or could rapidly transform to look larger or smaller or younger (Courbon and Tusques 1932; see also Ellis and Young 1990).

To summarize, recognition and categorization performance depends on rotations, size-scalings, translations, affine transformations, projective transformations, and topological transformations. Thus, performance in recognition and categorization tasks depends systematically on the amount of geometrical transformation, for almost all transformation groups of the *Erlanger Programm*. The only exception are mirror reflections, for which the amount of transformation is not defined, and therefore a systematic dependency cannot be expected. Given these similar patterns of performance in recognition and categorization tasks it seems reasonable and parsimonious to assume that both basic problems of recognition – the problem of shape constancy and the problem of class level recognition – rely on similar processing principles, and just involve different transformation groups. The recognition of individual objects after spatial transformations is mostly related to Euclidean transformations, like rotations, size-scalings and translations. In some cases, like a projection of a shape onto a slanted surface, also affine and projective transformations may be necessary. For object categorization the higher (deforming) transformation groups, especially nonlinear transformations, seem most important. Shape variability within basic and subordinate level categories usually involves space-curving (topological) transformations. However, simpler transformations may describe some of the shape variability within object categories, like affine stretching transformations for cups and containers (see Labov 1973), or size-scaling transformations if some category members simply differ in size. Thus, recognition and categorization can be coarsely assigned to involve different transformation groups in Klein's hierarchy, although there is no clear-cut mapping.

4 Integrative Transformational Framework of Recognition and Categorization

The general dependency on the amount of transformation for almost all transformation groups is difficult to reconcile with the notion of invariance, or with invariant properties. In principle, invariant property approaches predict that recognition performance does not depend on the amount of transformation, as invariants are by definition unaffected by transformations (e.g., Van Gool et al. 1994; Chen 1982, 1985; Palmer 1983, 1989, 1999). Hence, a systematic relation between recognition or categorization performance and the amount of geometrical transformation would not be predicted.[8] Thus, there are reasons to doubt that invariant property models equally account for object recognition and categorization. This extends also to other models which rely on abstract representations and predict that recognition and categorization performance is basically independent of geometrical transformations, like structural description models from the recognition literature (e.g., Biederman 1987; Hummel and Biederman 1992; Hummel and Stankiewicz 1998).

Most of the existing image-based models account for the systematic dependency of recognition and categorization performance on the amount of spatial transformations – even though many models do not involve explicit transformation processes to compensate for image transformations (e.g., Edelman 1998, 1999; Perrett et al. 1998; Riesenhuber and Poggio 1999, 2002; Wallis and Bülthoff 1999). However, there are two further important classes of findings which need to be accounted for by any model of object recognition (Graf 2006). One class relates to the notion that the visual system carries out analog transformation processes in object recognition, i.e., continuous or incremental transformation processes. Evidence for analog transformations comes from a study showing that rotation times in a sequential picture–picture matching task are sequentially additive (Bundesen et al. 1981; for review see Graf 2006). In other words, the transformations in the visual system seem to traverse intermediate points on the transformational path. Sequential additivity was demonstrated not only for rotations, but also for morph transformations in a categorization task (Graf 2002). Several further studies suggest analog transformation processes (Kourtzi and Shiffrar 2001; Georgopoulos 2000; Georgopoulos et al. 1989; Lurito et al. 1991; Wang et al. 1998), but the evidence is not yet conclusive (for discussion see Graf 2006).

Another important class of findings is related to congruency effects in object recognition, suggesting that object recognition involves the adjustment of a perceptual frame of reference, or coordinate system. For instance, misoriented objects are recognized better when a different object has been presented immediately before in the

[8] In a variation of this approach, Shepard (1994) claimed that the linearity of transformation time (in mental rotation tasks and apparent motion tasks) is an invariant.

same orientation (Gauthier and Tarr 1997; Graf et al. 2005; Jolicoeur 1990b, 1992; Tarr and Gauthier 1998). This orientation congruency effect appeared when the two objects were similar or dissimilar, when they belonged to the same or to different superordinate categories, and when the objects had the same or a different (horizontal vs. vertical) main axis of elongation (Graf et al. 2005). Congruency effects have been found also for size (e.g., Larsen and Bundesen 1978; Cave and Kosslyn 1989). These congruency effects suggest that recognition involves the adjustment of a perceptual reference frame or coordinate system, because performance is improved when the coordinate system is adjusted to the right orientation or size.

The existing image-based models cannot account for congruency effects in object recognition, because they are based on units that are simultaneously tuned to shape and orientation. Therefore, they do not predict a facilitation effect for the recognition of dissimilar shapes in the same orientation or size. Also current hybrid models which integrate image-based and structural representations (e.g., Edelman and Intrator 2000, 2001; Foster and Gilson 2002; Thoma et al. 2004) do not account for congruency effects. These models may account for congruency effects with similar parts, or similar structures, but not for dissimilar objects (see Graf 2006).

In order to integrate this large body of findings, Graf (2002, 2006) proposed a transformational framework of recognition and categorization. According to this transformational framework, the transformation groups of the *Erlanger Programm* describe *time-consuming (and error-prone) transformation processes* in the visual system. During recognition, differences in the spatial relation between memory representation and stimulus representation are compensated by a transformation of a perceptual coordinate system which brings memory and stimulus representations into correspondence. When both are aligned, a matching can be performed in a simpler way. Note that these transformations are not transformations of mental images as in mental rotation, but coordinate transformations (transformations of a perceptual coordinate system), and thus similar to transformations involved in visuomotor control tasks (Graf 2006; Salinas and Sejnowski 2001; Salinas and Abbott 2001). Transformations of the Euclidean similarity group (especially rotations, size-scalings and shifts in position) are usually sufficient to compensate for spatial transformations in object recognition (due to changes in the spatial relation between observer and object).

The transformational framework can be extended to account for categorization up to the basic level and compensate for shape differences – simply by allowing for nonlinear (deforming) transformations. Categorization is achieved by a deforming transformation which aligns memory and stimulus representations. For instance, Snoopy can be categorized as a dog by a topological (morphing) transformation which aligns the pictorial representation of the category dog and Snoopy's shape, until both can be matched. This approach provides an integrative framework of recognition and categorization up to the basic level, based on Klein's hierarchy of transformation groups. Recognition and categorization rely on similar processing principles, and differ mainly by involving different transformations (see Sect. 3). The transformational framework explains why recognition and categorization latencies depend in a systematic way on the amount of transformation which is

necessary for an alignment of memory representation and stimulus representation.[9] Moreover, the transformational framework is in accordance with findings showing that dynamic transformation processes are involved in categorization (Zaki and Homa 1999; see also Barsalou 1999), and with evidence for a transformational model of similarity (Hahn et al. 2003), including effects of the direction of a preceding morph transformation on subsequent similarity judgments (Hahn et al. 2009). Given the evidence that perceptual space is deformable and non-Euclidean (e.g., Hatfield 2003; Luneburg 1947; Suppes 1977; Watson 1978), the present framework suggests that physical space, perceptual space, and representational (categorical) space are endowed with a topological structure, leading to a unitary concept of space for these domains.

As morphing relies on an alignment of corresponding object parts or features, morphing can be regarded as an image-based structural alignment process. Thus, morphing may be an image-based instantiation of the structural alignment approach, which is prominent in the categorization literature (e.g., Gentner and Markman 1994, 1995; Goldstone and Medin 1994; Goldstone 1994a, 1994b, 1996; Markman 2001; Markman and Gentner 1993a, 1993b, 1997; Markman and Wisniewski 1997; Medin et al. 1993). The concept of image-based deforming transformations, or elastic matching, fits nicely with the idea of structured representations (Basri et al. 1998). Knowledge about the hierarchical organization of an object may also be important within an alignment approach, because this knowledge can guide the alignment process (Basri 1996), e.g., by facilitating the assignment of correspondent points or regions. This framework is compatible with evidence suggesting that object representations are structural or part-based (Biederman and Cooper 1991; Tversky and Hemenway 1984; Goldstone 1996; Newell et al. 2005; but see Cave and Kosslyn 1993; Murphy 1991), without having to assume abstract propositional representations.

According to the transformational framework, transformation processes are of primary importance – and not the search for invariant properties or features. This process-based view does not coincide with the invariants-based and static mathematical interpretation of the *Erlanger Programm* (as e.g., suggested by Niall 2000 or Cutting 1986, pp. 67–68). It seems more appropriate to assume a transformational framework of recognition and categorization, i.e., to focus on transformations and not on invariants. Nevertheless, invariants or features may be useful in a preselection process, or may play a role in solving the correspondence problem (e.g., Chen 2001; Carlsson 1999). Invariants may be involved in a fast feedforward sweep in visual processing which does not lead to a conscious percept, while conscious object perception seems to require recurrent processes (Lamme 2003; Lamme and Roelfsema 2000), potentially including transformational processes (Graf 2006).

[9] Not only the extent but also the type of topological distortion might be relevant (in analogy to the affine transformation group, cp. Wagemans et al. 1996). However, possible influences of the type of topological transformations are not in the focus of this work (for a discussion of the types of transformation see Bedford 2001).

To conclude, the transformational framework seems to be a highly parsimonious approach, because it is integrative in several different ways.

First, the transformational framework provides an integrative framework of recognition and categorization, based on Klein's hierarchy of transformation groups. Second, the transformational framework has the capacity to integrate image-based and structured (part-based) representations (see also Hahn et al. 2003), and can be regarded as an image-based extension of the structural alignment approach (e.g., Medin et al. 1993; Markman 2001). Third, according to the transformational framework, object recognition involves the adjustment of a perceptual coordinate system, i.e., involves coordinate transformations (Graf 2006). As coordinate transformations are fundamental also for visuomotor control, similar processing principles seem to be involved in object perception and perception for action (e.g., Graf 2006; Salinas and Abbott 2001; Salinas and Sejnowski 2001). This is compatible with the proposal that perception and action planning are coded in a common representational medium (e.g., Prinz 1990, 1997; Hommel et al. 2001). In accordance with this integrative approach, the recognition of manipulable objects can benefit from knowledge about typical motor interactions with the objects (Helbig et al. 2006). Finally, a framework of categorization based on pictorial representations and transformation processes is closely related to embodied approaches of cognition, like Larry Barsalou's (1999) framework of perceptual symbol systems, and embodied approaches to conceptual systems (Lakoff and Johnson 1999). The transformational framework does not invoke abstract propositional representations, but proposes that conceptual representations have a similar format as the perceptual input (see Graf 2002, 2006).

5 Open Questions and Outlook

The dependency of performance on the amount of transformation provides suggestive evidence that both object recognition and object categorization up to the basic level can be described by geometrical transformation processes. A process-based interpretation of the transformations of the *Erlanger Programm* offers the foundation for an integrative framework of object recognition and categorization. In any case, the *Erlanger Programm* provides a useful scheme to understand object shape, and to review the literature on object recognition and categorization.

Clearly, there are still open questions in the transformational framework. First, topological transformations are very powerful and can cross category boundaries. The question arises why, for instance, the dog template is aligned with a dog, but not with a cat, a cow, or a fish. Thus, constraints are necessary in order to avoid categorization errors. One important constraint is the transformational distance, which tends to be shorter within the same category than between categories. However, transformational distance alone might not be sufficient in all cases. Additional constraints may be provided by information about tolerable transformations (Bruce et al. 1991; see also Bruce 1994; Zaki and Homa 1999; see already

Murphy and Medin 1985; Landau 1994). The stored category exemplars may span some kind of space of tolerable topological transformations for each object category (Vernon 1952; Cootes et al. 1992; Baumberg and Hogg 1994; for a more detailed discussion see Graf 2002). In accordance, children at a certain age tend to overgeneralize and, for example, categorize many quadrupeds as dogs (e.g., Clark and Clark 1977; Waxman 1990).

A second problem is the so-called alignment paradox (e.g., Corballis 1988). It can be argued that an alignment through the shortest transformational path can only be achieved if the object is already identified or categorized. However, Ullman (1989, pp. 224–227) demonstrated that an alignment can be based on information which is available before identification or categorization (e.g., dominant orientation or anchor points), so that the paradox does not arise. This method can be used even for nonrigid transformations, if flexible objects are treated as locally rigid and planar. Further solutions to this correspondence problem have been proposed in the computer vision literature (e.g., Belongie et al. 2002; Carlsson 1999; Sclaroff 1997; Sclaroff and Liu 2001; Witkin et al. 1987; see also Ullman 1996). It should be noted that the problem is not fully solved yet. A potential resolution is that information necessary to perform an alignment is processed in a fast and unconscious feedforward sweep, while conscious recognition and categorization require recurrent processes (Lamme 2003; Lamme and Roelfsema 2000), like transformation processes.

Third, the neuronal implementation of a transformational framework of recognition and categorization is an open issue. It has been proposed that coordinate transformation processes in recognition and categorization are based on neuronal gain modulation (Graf 2006). Current approaches suggest that recognition and categorization are limited to the ventral visual stream (for reviews see e.g., Grill-Spector 2003; Grill-Spector and Sayres 2008; Malach et al. 2002). However, it seems possible that spatial transformation (and morphing) processes in recognition and categorization also involve the dorsal pathway, which is traditionally associated with spatial processing and coordinate transformations. There is suggestive evidence that the dorsal stream is involved in the recognition of objects that are rotated or size scaled (Eacott and Gaffan 1991; Faillenot et al. 1997, 1999; Gauthier et al. 2002; Harris et al. 2009; Kosslyn et al. 1994; Sugio et al. 1999; Vuilleumier et al. 2002; Warrington and Taylor 1973, 1978), and in the categorization of distorted dot patterns (Seger et al. 2000; Vogels et al. 2002).

Despite these open questions, the transformational framework seems promising due to its integrative potential. Moreover, if topological transformations are included into a framework of categorization, a number of further interesting issues can be tackled. First, deformations of objects due to nonrigid motion can be described by topological transformations. Many biological objects, including humans, deform when they move. In order to perceive an organism that is moving or adopting a new posture, the representation has to be updated, and deformations have to be compensated. Similarly, emotional facial expressions can be described by topological transformations (e.g., Knappmeyer et al. 2003). Second, the recognition of articulated objects can be covered by models that allow for deforming

transformations (e.g., Basri et al. 1998). Third, shape changes related to biological growth can be accounted for (Pittenger and Shaw 1975; Shaw and Pittenger 1977). Fourth, high-level adaptation phenomena, as reported in the face recognition literature (e.g., Leopold et al. 2001), can be described with deforming transformations. These high-level adaptation phenomena appear to involve deformations of the representational space. And fifth, the recognition of deformable objects, like for instance a rucksack, or cloths, etc., seems to require deformable transformations.

Author Note The work has been supported by a grant from the Max Planck Institute for Biological Cybernetics, Tübingen, and from the Max Planck Institute for Human and Cognitive Brain Sciences, Leipzig. I thank Christoph Dahl for substantial help in creating the 3D morph objects, some of which are depicted in Fig. 6.

References

Ashbridge E, Perrett DI (1998) Generalizing across object orientation and size. In: Walsh V, Kulikowski J (eds) Perceptual constancy. Why things look as they do. Cambridge University Press, Cambridge, pp 192–209

Ashby FG, Maddox WT (2005) Human category learning. Annu Rev Psychol 56:149–178

Barsalou LW (1999) Perceptual symbol systems. Behav Brain Sci 22:577–660

Basri R (1996) Recognition by prototypes. Int J Comput Vis 19:147–167

Basri R, Costa L, Geiger D, Jacobs D (1998) Determining the similarity of deformable shapes. Vis Res 38:2365–2385

Baumberg A, Hogg D (1994) Learning flexible models from image sequences. Proceedings of the Third European Conference on Computer Vision 1994. Springer, Berlin, pp 299–308

Bedford F (2001) Towards a general law of numerical/object identity. Curr Psychol Cogn 20:113–176

Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24:509–522

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115–147

Biederman I, Cooper EE (1991) Priming contour-deleted images: evidence for intermediate representations in visual object recognition. Cogn Psychol 23:393–419

Biederman I, Ju G (1988) Surface vs. edge-based determinants of visual recognition. Cogn Psychol 20:38–64

Bruce V (1994) Stability from variation: the case of face recognition. The M.D. Vernon memorial lecture. Q J Exp Psychol 47A:5–28

Bruce V, Burton M, Doyle T, Dench N (1989) Further experiments on the perception of growth in three dimensions. Percept Psychophys 46:528–536

Bruce V, Doyle T, Dench N, Burton M (1991) Remembering facial configurations. Cognition 38:109–144

Bundesen C, Larsen A (1975) Visual transformation of size. J Exp Psychol Hum Percept Perform 1:214–220

Bundesen C, Larsen A, Farrell JE (1981) Mental transformations of size and orientation. In: Long J, Baddeley A (eds) Attention and Performance, IX. Erlbaum, Hillsdale, NJ, pp 279–294

Carlsson S (1999) Order structure, correspondence and shape based categories. In Forsyth DA, Mundy JL, di Gesú V, Cipolla R (eds) Shape, contour and grouping in computer vision. Lecture Notes in Computer Science 1681. Springer, Berlin, pp 58–71

Cassirer E (1944) The concept of group and the theory of perception. Philos Phenomenol Res 5:1–35

Cave KR, Kosslyn SM (1989) Varieties of size-specific visual selection. J Exp Psychol Gen 118:148–164
Cave CB, Kosslyn SM (1993) The role of parts and spatial relations in object identification. Perception 22:229–248
Cave KR, Pinker S, Giorgi L, Thomas CE, Heller LM, Wolfe JM, Lin H (1994) The representation of location in visual images. Cogn Psychol 26:1–32
Chen L (1982) Topological structure in visual perception. Science 218:699–700
Chen L (1985) Topological structure in the perception of apparent motion. Perception 14:197–208
Chen L (2001) Perceptual organization: to reverse back the inverted (upside-down) question of feature binding. Vis Cogn 8:287–303
Chen L (2005) The topological approach to perceptual organization. Vis Cogn 12:553–637
Clark HH, Clark EV (1977) Psychology and language. Hartcourt Brace Jovanovich, New York
Cohen AL, Nosofsky RM (2000) An exemplar-retrieval model of speeded same-different judgments. J Exp Psychol Hum Percept Perform 26:1549–1569
Cooper EE, Biederman I (1993) Geon differences during object recognition are more salient than metric differences. Poster presented at the annual meeting of the Psychonomic Society. Washington DC
Cootes TF, Taylor CJ, Cooper DH, Graham J (1992) Training models of shape from sets of examples. Proceedings of the British Machine Vision Conference 1992. Springer, Berlin, pp 9–18
Corballis MC (1988) Recognition of disoriented shapes. Psychol Rev 95:115–123
Courbon P, Tusques J (1932) Illusions d'intermetamorphose et de charme. Annales Medico-Psychologiques 14:401–406
Cutting JE (1986) Perception with an eye for motion. MIT Univ Press, Cambridge, MA
Cutzu F, Edelman S (1996) Faithful representation of similarities among three-dimensional shapes in human vision. Proc Natl Acad Sci 93:12046–12050
Cutzu F, Edelman S (1998) Representation of object similarity in human vision: psychophysics and a computational model. Vis Res 38:2229–2257
Diamond R, Carey S (1986) Why faces are and are not special: an effect of expertise. J Exp Psychol Gen 115:107–117
Dill M, Edelman S (2001) Imperfect invariance to object translation in the discrimination of complex shapes. Perception 30:707–724
Dill M, Fahle M (1998) Limited translation invariance of human visual pattern recognition. Percept Psychophys 60:65–81
Dixon P, Just MA (1978) Normalization of irrelevant dimensions in stimulus comparisons. J Exp Psychol Hum Percept Perform 4:36–46
Dodwell PC (1983) The Lie transformation group model of visual perception. Percept Psychophys 34:1–16
Dürer A (1528) Vier Bücher von menschlicher proportion. Reprint on the basis of the original edition, 3rd edn (1996). Verlag Dr. Alfons Uhl, Nördlingen, Germany
Eacott MJ, Gaffan D (1991) The role of monkey inferior parietal cortex in visual discrimination of identity and orientation of shapes. Behav Brain Res 46:95–98
Edelman S (1995) Representation of similarity in three-dimensional object discrimination. Neural Comput 7:408–423
Edelman S (1997) Computational theories of object recognition. Trends Cogn Sci 1:296–304
Edelman S (1998) Representation is representation of similarities. Behav Brain Sci 21:449–498
Edelman S (1999) Representation and recognition in vision. MIT Univ Press, Cambridge, MA
Edelman S, Intrator N (2000) (Coarse coding of shape fragments) + (retinotopy) ≈ representation of structure. Spatial Vis 13:255–264
Edelman S, Intrator N (2001) A productive, systematic framework for the representation of visual structure. In: Lean TK, Dietterich TG, Tresp V (eds) Advances in neural information processing systems 13. MIT Univ Press, Cambridge, MA, pp 10–16
Ellis HD, Young AW (1990) Accounting for delusional misidentifications. Br J Psychiatry 157:239–248

Faillenot I, Toni I, Decety J, Grégoire M-C, Jeannerod M (1997) Visual pathways for object-oriented action and object recognition. Functional anatomy with PET. Cereb Cortex 7:77–85
Faillenot I, Decety J, Jeannerod M (1999) Human brain activity related to the perception of spatial features of objects. NeuroImage 10:114–124
Farah MJ (2000) The cognitive neuroscience of vision. Blackwell, Oxford, UK
Farah MJ, Hammond KM (1988) Mental rotation and orientation-invariant object recognition: dissociable processes. Cognition 29:29–46
Foster DH (1973) A hypothesis connecting visual pattern recognition and apparent motion. Kybernetik 13:151–154
Foster DH (1978) Visual apparent motion and the calculus of variations. In: Leeuwenberg ELJ, Buffart HFJM (eds) Formal theories of visual perception. Wiley, New York, pp 67–82
Foster DH, Gilson SJ (2002) Recognizing novel three-dimensional objects by summing signals from parts and views. Proc R Soc Lond B 269:1939–1947
Foster DH, Kahn JI (1985) Internal representations and operations in visual comparison of transformed patterns: effects of pattern point-inversion, positional symmetry, and separation. Biol Cybern 51:305–312
Gauthier I, Tarr MJ (1997) Orientation priming of novel shapes in the context of viewpoint-dependent recognition. Perception 26:51–73
Gauthier I, Hayward WG, Tarr MJ, Anderson AW, Skudlarski P, Gore JC (2002) BOLD activity during mental rotation and viewpoint-dependent object recognition. Neuron 34:161–171
Gentner D, Markman AB (1994) Structural alignment in comparison: no difference without similarity. Psychol Sci 5:152–158
Gentner D, Markman AB (1995) Similarity is like analogy. In: Cacciari C (ed) Similarity in language, thought, and perception. Brepols, Brussels, Belgium, pp 111–148
Georgopoulos AP (2000) Neural mechanisms of motor cognitive processes: functional MIT Univ Press and neurophysiological studies. In: Gazzaniga MS (ed) The new cognitive neurosciences. MIT Univ Press, Cambridge, MA, pp 525–538
Georgopoulos AP, Lurito JT, Petrides M, Schwartz AB, Massey JT (1989) Mental rotation of the neuronal population vector. Science 243:234–236
Gibson JJ (1950) The perception of the visual world. Houghton Mifflin, Boston
Goldstone RL (1994a) Similarity, interactive activation, and mapping. J Exp Psychol Learn Memory Cogn 20:3–28
Goldstone RL (1994b) The role of similarity in categorization: providing a groundwork. Cognition 52:125–157
Goldstone RL (1996) Alignment-based nonmonotonicities in similarity. J Exp Psychol Learn Memory Cogn 22:988–1001
Goldstone RL, Medin DL (1994) Time course of comparison. J Exp Psychol Learn Memory Cogn 20:29–50
Goodale MA, Milner AD (2004) Sight unseen. Oxford University Press, Oxford
Graf M (2002) Form, space and object. Geometrical transformations in object recognition and categorization. Wissenschaftlicher Verlag, Berlin
Graf M (2006) Coordinate transformations in object recognition. Psychol Bull 132:920–945
Graf M, Bülthoff HH (2003) Object shape in basic level categorisation. In Schmalhofer F, Young RM, Katz G (eds) Proceedings of the European Cognitive Science Conference. Lawrence Erlbaum, Mahwah, NJ, pp 390
Graf M, Kaping D, Bülthoff HH (2005) Orientation congruency effects for familiar objects: coordinate transformations in object recognition. Psychol Sci 16:214–221
Graf M, Bundesen C, Schneider WX (2009) Topological transformations in basic level object categorization. Manuscript submitted for publication
Grill-Spector K (2003) The neural basis of object perception. Curr Opin Neurobiol 13:1–8
Grill-Spector K, Sayres R (2008) Object recognition: insights from advances in fMRI methods. Curr Dir Psychol Sci 17:73–79
Hahn U, Chater N, Richardson LB (2003) Similarity as transformation. Cognition 87:1–32
Hahn U, Close J, Graf M (2009) Transformation direction influences shape-similarity judgments. Psychological Science, 20:447–454. DOI: 10.1111/j.1467-9280.2009.02310.x

Harris IM, Benito CT, Ruzzoli M, Miniussi C (2008) Effects of right parietal transcranial magnetic stimulation on object identification and orientation judgments. Journal of Cognitive Neuroscience, 20:916–926
Hatfield G (2003) Representation and constraints: the inverse problem and the structure of visual space. Acta Psychol 114:355–378
Hayward WG (2003) After the viewpoint debate: where next in object recognition? Trends Cogn Sci 7:425–427
Helbig HB, Graf M, Kiefer M (2006) The role of action representations in visual object recognition. Exp Brain Res 174:221–228
Homa D, Cross J, Cornell D, Goldman D, Shwartz S (1973) Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. J Exp Psychol 101:116–122
Hommel B, Müsseler J, Aschersleben G, Prinz W (2001) The theory of event coding (TEC): a framework for perception and action planning. Behav Brain Sci 24:849–937
Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. Psychol Rev 99:480–517
Hummel JE, Stankiewicz BJ (1998) Two roles for attention in shape perception: a structural description model of visual scrutiny. Vis Cogn 5:49–79
Ihmig K-N (1997) Cassirers Invariantentheorie der Erfahrung und seine Rezeption des "Erlanger Programms". Cassirer Forschungen, Band 2. Felix Meiner Verlag, Hamburg
Jolicoeur P (1985) The time to name disoriented natural objects. Memory Cogn 13:289–303
Jolicoeur P (1987) A size-congruency effect in memory for visual shape. Memory Cogn 15:531–543
Jolicoeur P (1988) Mental rotation and the identification of disoriented objects. Can J Psychol 42:461–478
Jolicoeur P (1990a) Identification of disoriented objects: a dual-systems theory. Mind Lang 5:387–410
Jolicoeur P (1990b) Orientation congruency effects on the identification of disoriented shapes. J Exp Psychol Hum Percept Perform 16:351–364
Jolicoeur P (1992) Orientation congruency effects in visual search. Can J Psychol 46:280–305
Jolicoeur P, Humphrey GK (1998) Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In Walsh V, Kulikowski J (eds) Perceptual constancy. Why things look as they do. Cambridge University Press, Cambridge, pp 69–123
Jolicoeur P, Gluck MA, Kosslyn SM (1984) Pictures and names: making the connection. Cogn Psychol 16:243–275
Jolicoeur P, Corballis MC, Lawson R (1998) The influence of perceived rotary motion on the recognition of rotated objects. Psychon Bull Rev 5:140–146
Kayaert G, Biederman I, Op de Beeck H, Vogels R (2005) Tuning for shape dimensions in macaque inferior temporal cortex. Eur J Neurosci 22:212–224
Klein F (1872/1893) Vergleichende Betrachtungen über neuere geometrische Forschungen. Mathematische Annalen 43:63–100
Knappmeyer B, Thornton IM, Bülthoff HH (2003) The use of facial motion and facial form during the processing of identity. Vis Res 43:1921–1936
Kosslyn SM (1994) Image and brain. MIT Univ Press, Cambridge, MA
Kosslyn SM, Alpert NM, Thompson WL, Chabris CF, Rauch SL, Anderson AK (1994) Identifying objects seen from different viewpoints. A PET investigation. Brain 117:1055–1071
Kourtzi Z, Shiffrar M (2001) Visual representation of malleable and rigid objects that deform as they rotate. J Exp Psychol Hum Percept Perform 27:335–355
Labov W (1973) The boundaries of words and their meanings. In: Bailey C-JN, Shuy RW (eds) New ways of analyzing variations in English. Georgetown University Press, Washington, D.C
Lakoff G (1987) Women, fire, and dangerous things. What categories reveal about the mind. University of Chicago Press, Chicago
Lakoff G, Johnson M (1999) Philosophy in the flesh: the embodied mind and Its challenge to western thought. Basic Books, New York, NY
Lamme VAF (2003) Why visual attention and awareness are different. Trends Cogn Sci 7:12–18

Lamme VAF, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci 23:571–579
Landau B (1994) Object shape, object name, and object kind: representation and development. In: Medin DL (ed) The psychology of learning and motivation 31. Academic Press, New York, pp 253–304
Larsen A, Bundesen C (1978) Size scaling in visual pattern recognition. J Exp Psychol Hum Percept Perform 4:1–20
Lawson R, Humphreys GW (1998) View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects. Percept Psychophys 60:1052–1066
Lawson R, Jolicoeur P (1998) The effects of plane rotation on the recognition of brief masked pictures of familiar objects. Memory Cogn 26:791–803
Lawson R, Jolicoeur P (1999) The effect of prior experience on recognition thresholds for plane-disoriented pictures of familiar objects. Memory Cogn 27:751–758
Lawson R, Humphreys GW, Jolicoeur P (2000) The combined effects of plane disorientation and foreshortening on picture naming: one manipulation are two? J Exp Psychol Hum Percept Perform 26:568–581
Leech R, Mareschal D, Cooper RP (2008) Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. Behavioral and Brain Sciences 31:357–378
Leopold DA, O'Toole AJ, Vetter T, Blanz V (2001) Prototype-referenced shape encoding revealed by high-level aftereffects. Nat Neurosci 4:89–94
Leyton M (1992) Symmetry, causality, mind. MIT Univ Press, Cambridge, MA
Lowe DG (1985) Perceptual organization and visual recognition. Kluwer, Boston, MA
Lowe DG (1987) Three-dimensional object recognition from single two-dimensional images. Artif Intell 31:355–395
Luneburg RK (1947) Mathematical analysis of binocular vision. Princeton University Press, Princeton, NJ
Lurito T, Georgakopoulos T, Georgopoulos AP (1991) Cognitive spatial-motor processes: 7. The making of movements at an angle from a stimulus direction: studies of motor cortical activity at the single cell and population levels. Exp Brain Res 87:562–580
Maddox WT, Ashby FG (1996) Perceptual separability, and the identification-speeded classification relationship. J Exp Psychol Hum Percept Perform 22:795–817
Malach R, Levy I, Hasson U (2002) The topography of high-order human object areas. Trends Cogn Sci 6:176–184
Mark LS, Todd JT (1985) Describing perceptual information about human growth in terms of geometric invariants. Percept Psychophys 37:249–256
Markman AB (2001) Structural alignment, similarity, and the internal structure of category representations. In: Hahn U, Ramscar M (eds) Similarity and categorization. Oxford University Press, Oxford, pp 109–130
Markman AB, Gentner D (1993a) Splitting the differences: a structural alignment view of similarity. J Memory Lang 32:517–535
Markman AB, Gentner D (1993b) Structural alignment during similarity comparisons. Cogn Psychol 25:431–467
Markman AB, Gentner D (1997) The effects of alignability on memory. Psychol Sci 8:363–367
Markman AB, Wisniewski EJ (1997) Similar and different: the differentiation of basic-level categories. J Exp Psychol Learn Memory Cogn 23:54–70
Medin DL, Goldstone RL, Gentner D (1993) Respects for similarity. Psychol Rev 100:254–278
Michaels CF, Carello C (1981) Direct perception. Prenctice-Hall, Englewood Cliffs, NJ
Milliken B, Jolicoeur P (1992) Size effects in visual recognition memory are determined by perceived size. Memory Cogn 20:83–95
Milner AD, Goodale MA (1995) The visual brain in action. Oxford University Press, Oxford, England
Murphy GL (1991) Parts in object concepts: experiments with artificial categories. Memory Cogn 19:423–438

Murphy GL (2002) The big book of concepts. MIT Univ Press, Cambridge, MA
Murphy GL, Medin DL (1985) The role of theories in conceptual coherence. Psychol Rev 92:289–316
Nazir TA, O'Regan JK (1990) Some results on translation invariance in the human visual system. Spat Vis 5:81–100
Newell FN, Sheppard DM, Edelman S, Shapiro KL (2005) The interaction of shape- and location-based priming in object categorisation: evidence for a hybrid "what+where" representation stage. Vis Res 45:2065–2080
Niall KK (1992) Projective invariance and the kinetic depth effect. Acta Psychol 81:127–168
Niall KK (2000) Some plane truths about pictures: notes on Wagemans, Lamote, and van Gool (1997). Spat Vis 13:1–24
Niall KK, Macnamara J (1990) Projective invariance and picture perception. Perception 19:637–660
Nosofsky RM (1986) Attention, similarity, and the identification-categorization-relationship. J Exp Psychol Gen 115:39–57
Palmer SE (1983) The psychology of perceptual organization: a transformational approach. In: Beck J, Hope B, Rosenfeld A (eds) Human and machine vision. Academic Press, New York, pp 269–339
Palmer SE (1989) Reference frames in the perception of shape and orientation. In: Shepp BE, Ballesteros S (eds) Object perception: structure and process. Erlbaum, Hillsdale, NJ, pp 121–163
Palmer SE (1999) Vision science. Photons to phenomenology. MIT Univ Press, Cambridge, MA
Palmer SE, Rosch E, Chase P (1981) Canonical perspective and the perception of objects. In: Long J, Baddeley A (eds) Attention and performance IX. Erlbaum, Hillsdale, NJ, pp 135–151
Palmeri TJ, Gauthier I (2004) Visual object understanding. Nat Rev Neurosci 5:1–13
Perrett DI, Oram WM, Ashbridge E (1998) Evidence accumulation in cell populations responsive to faces: an account of generalization of recognition without mental transformations. Cognition 67:111–145
Pittenger JB, Shaw RE (1975) Aging faces as viscal-elastic events: implications for a theory of nonrigid shape perception. J Exp Psychol Hum Percept Perform 1:374–382
Pittenger JB, Shaw RE, Mark LS (1979) Perceptual information for the age level of faces as a higher order invariant of growth. J Exp Psychol Hum Percept Perform 5:478–493
Pitts W, McCulloch WS (1947) How we know universals: the perception of auditory and visual forms. Bull Math Biophys 9:127–147
Pizlo Z (1994) A theory of shape constancy based on perspective invariants. Vis Res 34:1637–1658
Posner MI, Keele S (1968) On the genesis of abstract ideas. J Exp Psychol 77:353–363
Posner MI, Keele S (1970) Retention of abstract ideas. J Exp Psychol 83:304–308
Posner MI, Goldsmith R, Welton KE (1967) Perceived distance and the classification of distorted patterns. J Exp Psychol 73:28–38
Prinz W (1990) A common coding approach to perception and action. In: Neumann O, Prinz W (eds) Relationships between perception and action: current approaches. Springer, Berlin, New York, pp 167–201
Prinz W (1997) Perception and action planning. Eur J Cogn Psychol 9:129–154
Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025
Riesenhuber M, Poggio T (2002) Neural mechanisms of object recognition. Curr Opin Neurobiol 12:162–168
Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P (1976) Basic objects in natural categories. Cogn Psychol 8:382–439
Salinas E, Abbott LF (2001) Coordinate transformations in the visual system: how to generate gain fields and what to compute with them. In Nicolelis MAL (ed) Advances in neural population coding. Progress in brain research, vol 130. Elsevier, Amsterdam, pp 175–190

Salinas E, Sejnowski TJ (2001) Gain modulation in the central nervous system: where behavior, neurophysiology, and computation meet. Neuroscientist 7:430–440
Schweinberger SR, Burton AM, Kelly SW (1999) Asymmetric dependencies in perceiving identity and emotion: experiments with morphed faces. Percept Psychophys 61:1102–1115
Schyns PG (1997) Categories and concepts: a bi-directional framework for categorization. Trends Cogn Sci 1:183–189
Sclaroff S (1997) Deformable prototypes for encoding shape categories in image databases. Pattern Recogn 30:627–642
Sclaroff S, Liu L (2001) Deformable shape detection and description via model-based region grouping. IEEE Trans Pattern Anal Mach Intell 23:475–489
Seger CA, Poldrack RA, Prabhakaran V, Zhao M, Glover G, Gabrieli JDE (2000) Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. Neuropsychologia 38:1316–1324
Shaw R, Pittenger J (1977) Perceiving the face of change in changing faces: implications for a theory of object perception. In Shaw R, Bransford J (eds) Perceiving, acting, and knowing. Toward an ecological psychology. Erlbaum, Hillsdale, NJ pp. 103–132
Shepard RN (1994) Perceptual-cognitive universals as reflections of the world. Psychon Bull Rev 1:2–28
Srinivas K (1993) Perceptual specificity in nonverbal priming. J Exp Psychol Learn Memory Cogn 19:582–602
Sugio T, Inui T, Matsuo K, Matsuzawa M, Glover GH, Nakai T (1999) The role of the posterior parietal cortex in human object recognition: a functional magnetic resonance imaging study. Neurosci Lett 276:45–48
Suppes P (1977) Is visual space Euclidean? Synthese 35:397–421
Tarr MJ (2003) Visual object recognition: can a single mechanism suffice? In Peterson MA, Rhodes G (eds) Perception of faces, objects, and scenes. Analytic and holistic processes. Oxford University Press, Oxford, England, pp 177–211
Tarr MJ, Gauthier I (1998) Do viewpoint-dependent mechanisms generalize across members of a class? Cognition 67:71–109
Tarr MJ, Williams P, Hayward WG, Gauthier I (1998) Three-dimensional object recognition is viewpoint-dependent. Nat Neurosci 1:275–277
Thoma V, Hummel JE, Davidoff J (2004) Evidence for holistic representations of ignored images and analytic representations of attended images. J Exp Psychol Hum Percept Perform 30:257–267
Thompson D'AW (1917) On growth and form, 2nd edn. (1942). Cambridge University Press, Cambridge
Todd JT, Chen L, Norman JF (1998) On the relative salience of Euclidean, affine, and topological structure for 3-D form discrimination. Perception 27:273–282
Tversky B, Hemenway K (1984) Objects, parts, and categories. J Exp Psychol Gen 113:169–193
Ullman S (1989) Aligning pictorial descriptions: an approach to object recognition. Cognition 32:193–254
Ullman S (1996) High-level vision. Object recognition and visual cognition. MIT Univ Press, Cambridge, MA
Ullman S (2007) Object recognition and segmentation by a fragment-based hierarchy. Trends Cogn Sci 11:58–64
Ungerleider LG, Haxby JV (1994) 'What' and 'where' in the human brain. Curr Opin Neurobiol 4:157–165
Ungerleider LG, Mishkin M (1982) Two cortical visual systems. In: Ingle DJ, Goodale MA, Mansfield RJW (eds) Analysis of visual behavior. MIT Univ Press, Cambridge, MA, pp 549–586
Van Gool LJ, Moons T, Pauwels E, Wagemans J (1994) Invariance from the Euclidean geometer's perspective. Perception 23:547–561
Vernon MD (1952) A further study of visual perception. Cambridge University Press, London
Vogels R, Sary G, Dupont P, Orban GA (2002) Human brain regions involved in visual categorization. NeuroImage 16:401–414

Vuilleumier P, Henson RN, Driver J, Dolan RJ (2002) Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. Nat Neurosci 5:491–499
Wagemans J, Van Gool L, Lamote C (1996) The visual systems measurement of invariants need not itself be invariant. Psychol Sci 7:232–236
Wagemans J, Lamote C, Van Gool L (1997) Shape equivalence under perspective and projective transformations. Psychon Bull Rev 4:248–253
Wallis G, Bülthoff H (1999) Learning to recognize objects. Trends Cogn Sci 3:22–31
Wang G, Tanifuji M, Tanaka K (1998) Funcitional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. Neurosci Res 32:33–46
Warren WH, Shaw RE (1985) Events and encounters as units of analysis for ecological psychology. In Warren WH, Shaw RE (eds) Persistence and change. Proceedings of the first international conference on event perception. Erlbaum, Hillsdale, NJ, pp 1–27
Warrington EK, Taylor AM (1973) The contribution of the right parietal lobe to object recognition. Cortex 9:152–164
Warrington EK, Taylor AM (1978) Two categorical stages of object recognition. Perception 7:695–705
Watson AB (1978) A Riemann geometric explanation of the visual illusions and figural aftereffects. In: Leeuwenberg E, Buffart H (eds) Formal theories of visual perception. Wiley, New York, NY, pp 139–169
Waxman SR (1990) Linguistic biases and the establishment of conceptual hierarchies: evidence from preschool children. Cogn Dev 5:123–150
Willems B, Wagemans J (2001) Matching multicomponent objects from different viewpoints: mental rotation as normalization? J Exp Psychol Hum Percept Perform 27:1090–1115
Witkin A, Terzopoulos D, Kass M (1987) Signal matching through scale space. Int J Comput Vis 2:133–144
Young AW, Rowland D, Calder AJ, Etcoff NL, Seth A, Perrett DI (1997) Facial expression megamix: tests of dimensional and category accounts of emotion recognition. Cognition 63:271–313
Zaki SR, Homa D (1999) Concepts and transformational knowledge. Cogn Psychol 39:69–115

Comparison

Robert L. Goldstone, Sam Day, and Ji Y. Son

Abstract The process of comparison plays a critical role in problem solving, judgment, decision making, categorization, and cognition, broadly construed. In turn, determination of similarities and differences plays a critical role for comparison. In this chapter, we describe important classes of formal models of similarity and comparison: geometric, featural, alignment-based, and transformational. We also consider the question of whether similarity is too flexible to provide a stable ground for cognition, and conversely, whether it is insufficiently flexible to account for the sophistication of cognition. Both similarity assessments and comparison are argued to provide valuable general-purpose cognitive strategies.

1 Introduction

It might not be immediately clear why the topic of comparison warrants a whole chapter in a book on human thinking. Of course, we are often required to make decisions that involve comparing two or more alternatives and assessing their relative value. Which car should I buy? Which job is more suited to my long-term goals? Would I rather have the soup or the salad? But in the grand scheme of human cognition, it might seem that such processes could be relegated to a subheading in a chapter on decision making.

In fact, comparison is one of the most integral components of human thought. Along with the related construct of *similarity*, comparison plays a crucial role in almost everything that we do. Furthermore, comparison itself is a powerful cognitive tool – in addition to its supporting role in other mental processes, research has demonstrated that the simple act of comparing two things can produce important changes in our knowledge.

R.L. Goldstone (✉), S. Day, and J.Y. Son
Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, 47405, USA
e-mail: rgoldsto@indiana.edu

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_7,

One primary function of comparison is simply to assess the similarity of two things. To understand why this is such an important part of cognition, consider the variety of processes that are hypothesized to use similarity as an input. In models of memory, recognition and reminding have been argued to rely on the similarity between a stimulus and a long-term representation (Hintzman 1986; Shiffrin and Steyvers 1997). Models of categorization have proposed that new examples are classified based on their similarity to other category members (Medin and Shaffer 1978; Nosofsky 1984), or to a prototype of a category (Reed 1972). When making inferences about unknown properties, people often appear to rely on their knowledge about other similar entities and situations to make reasonable predictions (Osherson et al. 1990; Shepard 1987), and people are very likely to look to similar situations from their past when understanding and solving new problems (Holyoak and Koh 1987; Ross 1989). Thus, it is a rare moment in our lives when comparison and similarity do *not* seem to play a role.

However, comparison does more than simply assess existing representations – it can also affect our understanding of the things that are being compared. For example, research in decision making has shown that people's judgments and preferences may vary significantly based on the particular comparisons that are made (Huber et al. 1982; Simonson 1989). More direct evidence comes from Medin et al. (1993), who found that participants interpreted the features of an item differently when it had been compared to different alternatives. For example, in the top row of Fig. 1, when the ambiguous object B is compared to A, participants often write that a similarity between the pair is that both shapes have three prongs. However, when B is paired with C instead, participants often write that a similarity between the pair is that they both possess four prongs, and a difference is that one of B's prongs is warped or stunted. In other words, the comparison process seems to determine the content of our representations.

Importantly, these representational changes often appear to be of a very beneficial kind: comparison can allow an individual to look past simple "surface" features, and to focus instead on potentially more meaningful structural commonalities and differences. For example, (Gentner and Namy 1999; Namy and Gentner 2002) found that comparing two objects allowed young children to overcome their strong bias for perceptual similarity, and to group objects instead on common taxonomic membership. Even more impressively, research has shown that a previous comparison can change the way that people interpret *new* situations. When people compare two cases that share the same underlying principle, they are far more likely to recognize new cases where that principle is applicable (e.g., Gick and Holyoak 1983; Gentner et al. 2003). This improvement does not occur if the two cases are evaluated independently, without comparison (see Gentner's chapter on analogy in this book for a more detailed account of these kinds of effects). Even comparing situations that have slightly different underlying structures can be very beneficial, because it tends to highlight those structural differences (so-called "near miss" cases; Winston 1975).

Comparison therefore provides an invaluable tool for learning, allowing people to see how two things are alike and different, and to see important features of each

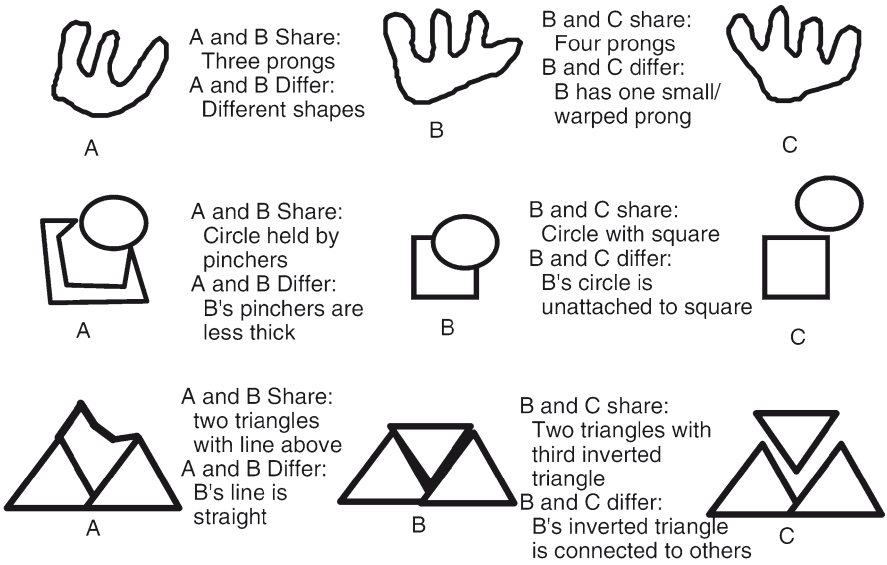

Fig. 1 Examples of stimuli from Medin et al. (1993). Subjects were asked to describe features that were shared and different between pairs of objects. The middle objects labeled *B* are ambiguous, and tend to be interpreted in a manner that is consistent with the objects (*A* or *C*) with which they are paired. When determining both common and distinctive features, people apparently first interpret objects so as to make them more comparable

case that might otherwise have been overlooked. This helps to explain why educational assignments that ask a student to "compare and contrast" are such a powerful tool (i.e., Bransford and Schwartz 1999), and makes it that much more puzzling that these types of assignments seem to have fallen out of favor in recent years.

2 Models of Similarity

Given the cognitive importance of comparison, it is understandable that there have been several attempts to formalize the comparison process. The formal treatments frequently center on the question of what makes things seem similar to people. One of the prominent goals of comparison is to determine how, and in what ways, two objects, scenes, or entities are similar to one another.

The formal treatments of similarity simultaneously provide theoretical accounts of similarity and describe how it can be empirically measured (Hahn 2003). These models have had a profound practical impact in statistics, automatic pattern recognition by machines, data mining, and marketing (e.g., online stores can provide "people similar to you liked the following other items…"). Our brief survey is organized in terms of the following models: geometric, feature-based, alignment-based, and transformational. It should be noted that although these models are laudable for

their quantitative predictions, they also bypass the important issue of what counts as a psychologically significant description of an object in the first place. These models adopt a philosophy of "You tell me what the features/dimensions/attributes/relations of an object are, and I will tell you how they are integrated together to come up with an impression of similarity." In fact, this attitude downplays the hard cognitive work in comparison that involves coming up with these descriptions in the first place (Goldstone et al. 1997; Hofstadter 1997; Shanon 1988). To be complete cognitive models, at the very least the models described below need to be supplemented by perceptual and conceptual processes that provide input descriptions. Furthermore, even this division of cognitive labor into representational and comparison processes has been questioned. As mentioned earlier, these two cognitive acts cannot be so cleanly separated because the very act of comparison alters one's descriptions of the compared objects.

2.1 Geometric Models and Multidimensional Scaling

Geometric models of similarity have been among the most influential approaches to analyzing similarity (Carroll and Wish 1974; Torgerson 1965). These approaches are exemplified by nonmetric multidimensional scaling (MDS) models (Shepard 1962a, 1962b). MDS models represent similarity relations between entities in terms of a geometric model that consists of a set of points embedded in a dimensionally organized metric space. The input to MDS routines may be similarity judgments, dissimilarity judgments, confusion matrices, correlation coefficients, joint probabilities, or any other measure of pairwise proximity. The output of an MDS routine is a geometric model of the data, with each object of the data set represented as a point in an *n*-dimensional space. The similarity between a pair of objects is taken to be inversely related to the distance between two objects' points in the space. In MDS, the distance between points *i* and *j* is typically computed by:

$$\text{dissimilarity}(i,j) = \left[\sum_{k=1}^{n} \left|X_{ik} - X_{jk}\right|^{r}\right]^{\frac{1}{r}}, \tag{1}$$

where n is the number of dimensions, X_{ik} is the value of dimension k for item i, and r is a parameter that allows different spatial metrics to be used. With $r = 2$, a standard Euclidean notion of distance is invoked, whereby the distance between two points is the length of the straight line connecting the points. If $r = 1$, then distance involves a city-block metric where the distance between two points is the sum of their distances on each dimension ("short-cut" diagonal paths are not allowed to directly connect points differing on more than one dimension). A Euclidean metric often provides a better fit to empirical data when the stimuli being compared are composed of integral, perceptually fused dimensions such as the brightness and saturation of a color. Conversely, a city-block metric is often appropriate for psychologically separated dimensions such as brightness and size (Attneave 1950).

A study by Smith et al. (1973) illustrates a classic use of MDS. They obtained similarity ratings from subjects on many pairs of birds. Submitting these pairwise similarity ratings to MDS analysis, they obtained the results shown in Fig. 2a (Fig. 2b shows a second analysis involving animals more generally). The MDS algorithm produced this geometric representation by positioning the birds in a two-dimensional space such that birds that are rated as being highly similar are very close to each other in the space. One of the main applications of MDS is to determine the underlying dimensions comprising the set of compared objects. Once the points are positioned in a way that faithfully mirrors the subjectively obtained similarities, it is often possible to give interpretations to the axes, or to rotations of the axes. Assigning subjective interpretations to the geometric model's axes, the experimenters suggested that birds were represented in terms of their values on dimensions such as "ferocity" and "size." It is important to note that the proper psychological interpretation of a geometric representation of objects is not necessarily in terms of its Cartesian axes. In some domains, such as musical pitches, the best interpretation of objects may be in terms of their polar coordinates of angle and length (Shepard 1982). Recent work has extended geometric representations still further, representing patterns of similarities by generalized, nonlinear manifolds (Tenenbaum et al. 2000).

Another use of MDS is to create quantitative representations that can be used in mathematical and computational models of cognitive processes. Numeric representations, namely coordinates in a psychological space, can be derived for stories, pictures, sounds, words, or any other stimuli for which one can obtain subjective similarity data. Once constructed, these numeric representations can be used to

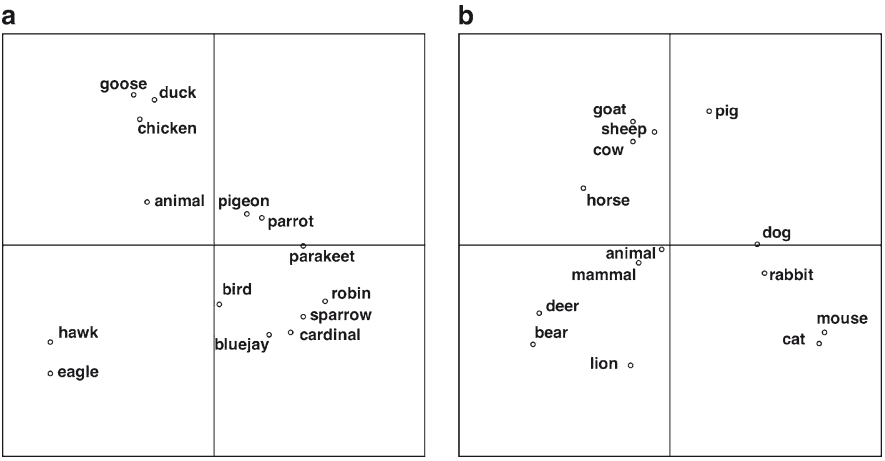

Fig. 2 Two multidimensional scaling (MDS) solutions for sets of birds (**a**) and animals (**b**). The distances between the animals in the space reflect their psychological dissimilarity. Once an MDS solution has been made, psychological interpretations for the dimensions may be possible. In these solutions, the *horizontal* and *vertical* dimensions may represent size and domesticity, respectively (Reprinted from Rips et al. 1973, by permission)

predict people's categorization accuracy, memory performance, or learning speed. MDS models have been successful in expressing cognitive structures in stimulus domains as far removed as animals (Smith et al. 1974), Rorschach ink blots (Osterholm et al. 1985), chess positions (Horgan et al. 1989), and air flight scenarios (Schvaneveldt et al. 1985). Many objects, situations, and concepts seem to be psychologically structured in terms of dimensions, and a geometric interpretation of the dimensional organization captures a substantial amount of that structure.

Obtaining all pairwise similarity ratings among a large set of objects is, experimentally speaking, effortful. For N objects, N^2 ratings are required as input to a standard MDS algorithm. However, geometric models of similarity have received a recent boost from automated techniques for analyzing large corpora of text. A computational approach to word meaning that has received considerable recent attention has been to base word meanings solely on the patterns of cooccurrence between a large number of words in an extremely large text corpus (Burgess and Lund 2000; Griffiths et al. 2007; Landauer and Dumais 1997). Mathematical techniques are used to create vector encodings of words that efficiently capture their cooccurrences. If two words, such as "cocoon" and "butterfly" frequently cooccur in an encyclopedia or enter into similar patterns of cooccurrence with other words, then their vector representations will be highly similar. The meaning of a word, its vector in a high dimensional space, is completely based on the contextual similarity of words to other words. Within this high dimensional space, Landauer and Dumais (1997) conceive of similarity as the cosine of the angle between two words rather than their distance. With these new techniques, it is now possible to create geometric spaces with tens of thousands of words.

2.2 Featural Models

In 1977, Amos Tversky brought into prominence what would become the main contender to geometric models of similarity in psychology. The reason given for proposing a feature-based model was that subjective assessments of similarity did not always satisfy the assumptions of geometric models of similarity:

Minimality: $D(A,B) \geq D(A,A) = 0$

Symmetry: $D(A,B) = D(B,A)$

The Triangle Inequality: $D(A,B) + D(B,C) \geq D(A,C)$

where $D(A,B)$ is interpreted as the dissimilarity between items A and B.

Violations of all three assumptions have been empirically obtained (Polk et al. 2002; Tversky 1977; Tversky and Gati 1982; Tversky and Hutchinson 1986). In light of the above potential problems for geometric representations, Tversky (1977) proposed to characterize similarity in terms of a feature-matching process based on weighting common and distinctive features. In this model, entities are represented as a collection of features and similarity is computed by:

$$S(A,B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A).$$

The similarity of A to B is expressed as a linear combination of the measure of the common and distinctive features. The term $(A \cap B)$ represents the features that items A and B have in common. $(A - B)$ represents the features that A has but B does not. $(B - A)$ represents the features of B that are not in A. θ, α and β are weights for the common and distinctive components. Common features as compared to distinctive features, are given relatively more weight for verbal as opposed to pictorial stimuli (Gati and Tversky 1984), for coherent as opposed to noncoherent stimuli (Ritov et al. 1990), for similarity as opposed to difference judgments (Tversky 1977), and for entities with a large number of distinctive as opposed to common features (Gati and Tversky 1984). There are no restrictions on what may constitute a feature. A feature may be any property, characteristic or aspect of a stimulus. Features may be concrete or abstract (i.e., "symmetric" or "beautiful").

The Contrast Model predicts asymmetric similarity because α is not constrained to equal β and $f(A-B)$ may not equal $f(B - A)$. North Korea is predicted to be more similar to Red China than vice versa if Red China has more salient distinctive features than North Korea, and α is greater than β. The Contrast Model can also account for nonmirroring between similarity and difference judgments. The common features term $(A \cap B)$ is hypothesized to receive more weight in similarity than difference judgments; the distinctive features term receives relatively more weight in difference judgments. As a result, certain pairs of stimuli may be perceived as simultaneously being more similar to and more different from each other, compared to other pairs (Tversky 1977). Sixty-seven percent of a group of subjects selected West Germany and East Germany as more similar to each other than Ceylon and Nepal. Seventy percent of subjects also selected West Germany and East Germany as more different from each other than Ceylon and Nepal. According to Tversky, East and West Germany have more common and more distinctive features than Ceylon and Nepal.

A number of models are similar to the Contrast model in basing similarity on features and in using some combination of the $(A \cap B)$, $(A - B)$, and $(B - A)$ components. Sjoberg (1972) proposes that similarity is defined as $f(A \cap B)/f(A \cup B)$. Eisler and Ekman (1959) claim that similarity is proportional to $f(A \cap B)/(f(A) + f(B))$. Bush and Mosteller (1951) defines similarity as $f(A \cap B)/f(A)$. These three models can all be considered specializations of the general equation $f(A \cap B)/[f(A \cap B)+\alpha f(A - B)+\beta f(B - A)]$. As such, they differ from the Contrast model by applying a ratio function as opposed to a linear contrast of common and distinctive features.

The fundamental premise of the Contrast Model, that entities can be described in terms of constituent features, is a powerful idea in cognitive psychology. Featural analyses have proliferated in domains of speech perception (Jakobson et al. 1963), pattern recognition (Neisser 1967; Treisman 1986), perception physiology (Hubel and Wiesel 1968), semantic content (Katz and Fodor 1963), and categorization (Medin and Shaffer 1978). Neural network representations are often based on features, with entities being broken down into a vector of ones and zeros, where each bit refers to a feature or "microfeature." Similarity plays a crucial role in many connectionist theories of generalization, concept formation, and learning. The notion of dissimilarity used in these systems is typically the fairly simple function

"Hamming distance." The Hamming distance between two strings is simply their city-block distance; that is, it is their $(A - B) + (B - A)$ term. "1 0 0 1 1" and "1 1 1 1 1" would have a Hamming distance of 2 because they differ on two bits. Occasionally, more sophisticated measures of similarity in neural networks normalize dissimilarities by string length. Normalized Hamming distance functions can be expressed by $[(A - B) + (B - A)]/[f(A \cap B)]$.

2.3 Similarities Between Geometric and Feature-Based Models

While MDS and featural models are often analyzed in terms of their differences, they also share a number of similarities. Recent progress has been made on combining both representations into a single model, using Bayesian statistics to determine whether a given source of variation is more efficiently represented as a feature or dimension (Navarro and Lee 2004). Tversky and Gati (1982) described methods of translating continuous dimensions into featural representations. Dimensions that are sensibly described as being more or less (e.g., loud is more sound than soft, bright is more light than dim, and large is more size than small) can be represented by sequences of nested feature sets. That is, the features of B include a subset of A's features whenever B is louder, brighter, or larger than A. Alternatively, for qualitative attributes like shape or hue (red is not subjectively "more" than blue), dimensions can be represented by chains of features such that if B is between A and C on the dimension, then $(A \cap B) \supset (A \cap C)$ and $(B \cap C) \supset (A \cap C)$. For example, if orange lies between red and yellow on the hue dimension, then this can be featurally represented by orange sharing features with both red and yellow, features that red and yellow do not share between themselves.

An important attribute of MDS models is that they create *postulated* representations, namely dimensions, that explain the systematicities present in a set of similarity data. This is a classic use of abductive reasoning; dimensional representations are hypothesized that, if they were to exist, would give rise to the obtained similarity data. Other computational techniques share with MDS the goal of discovering the underlying descriptions for items of interest, but create featural rather than dimensional representations. Hierarchical Cluster Analysis, like MDS, takes pairwise proximity data as input. Rather than output a geometric space with objects as points, Hierarchical Cluster Analysis outputs an inverted-tree diagram, with items at the root-level connected with branches. The smaller the branching distance between two items, the more similar they are. Just as the dimensional axes of MDS solutions are given subjective interpretations, the branches are also given interpretations. For example, in Shepard's (1972) analysis of speech sounds, one branch is interpreted as voiced phonemes while another branch contains the unvoiced phonemes. In additive cluster analysis (Shepard and Arabie 1979) similarity data is transformed into a set of overlapping item clusters. Items that are highly similar will tend to belong to the same clusters. Each cluster can be considered as a feature. Recent progress has been made on efficient and mathematically principled models that find

such featural representations for large databases (Lee 2002; Navarro and Griffiths 2007; Tenenbaum 1996).

Another commonality between geometric and featural representations, one that motivates the next major class of similarity models that we consider, is that both use relatively unstructured representations. Entities are structured as sets of features or dimensions with no relations between these attributes. Entities such as stories, sentences, natural objects, words, scientific theories, landscapes, and faces are not simply a "grab bag" of attributes. Two kinds of structure seem particularly important: propositional and hierarchical. A proposition is an assertion about the relation between informational entities (Palmer 1975). For example, relations in a visual domain might include *Above*, *Near*, *Right, Inside*, and *Larger-than* that take informational entities as arguments. The informational entities might include features such as *square*, and values on dimensions such as *3 in.* Propositions are defined as the smallest unit of knowledge that can stand as a separate assertion and have a truth value. The order of the arguments in the predicate is critical. For example, *above (Triangle, Circle)* does not represent the same fact as *Above (Circle, Triangle)*. Hierarchical representations involve entities that are embedded in one another. Hierarchical representations are required to represent the fact that *X* is *part of Y* or that *X* is a *kind of Y*. For example, in Collins and Quillian's (1969) propositional networks, labeled links ("Is-a" links) stand for the hierarchical relation between *Canary* and *Bird*.

Geometric and featural accounts of similarity fall short of a truly general capacity to handle structured inputs. Figure 3 shows an example of the need for structured representations . Using these materials 20 undergraduates were shown triads consisting of *A*, *B*, and *T*, and we asked them to decide whether Scene *A* or *B* was more similar to *T*. The strong tendency to choose *A* over *B* in the first panel suggests that the feature "square" influences similarity. Other choices indicated that

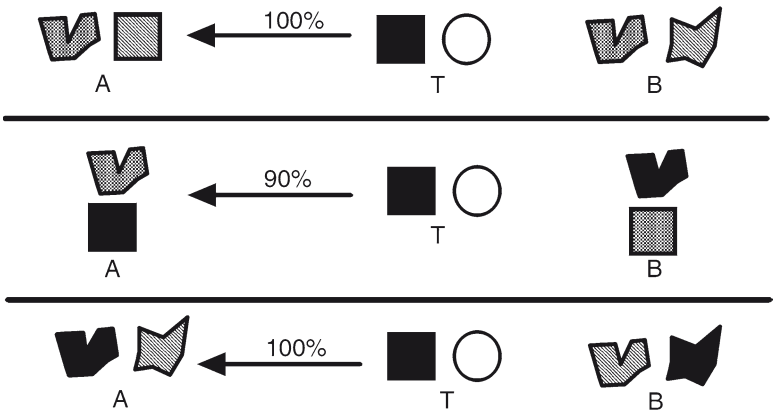

Fig. 3 The sets of objects *T* are typically judged to be more similar to the objects in the *A* sets than the *B* sets. These judgments show that people pay attention to more than just simple properties like "black" or "square" when comparing scenes

subjects also based similarity judgments on the spatial locations and shadings of objects as well as their shapes.

However, it is not sufficient to represent the left-most object of *T* as {Left, Square, Black} and base similarity on the number of shared and distinctive features. In the second panel, *A* is again judged to be more similar to *T* than is *B*. Both *A* and *B* have the features "Black" and "Square." The only difference is that for *A* and *T*, but not *B*, the "Black" and "Square" features belong to the same object. This is only compatible with feature set representations if we include the possibility of *conjunctive features* in addition to *simple features* such as "Black" and "Square" (Gluck 1991; Hayes-Roth and Hayes-Roth 1977). By including the conjunctive feature "Black-Square," possessed by both *T* and *A*, we can explain, using feature sets, why *T* is more similar to *A* than *B*. The third panel demonstrates the need for a "Black-Left" feature, and other data indicates a need for a "Square-Left" feature. Altogether, if we wish to explain similarity judgments that people make we need a feature set representation that includes six features (three simple and three complex) to represent the square of *T*.

However, there are two objects in *T*, bringing the total number of features required to at least two times the six features required for one object. The number of features required increases still further if we include feature-triplets such as "Left-Black-Square." In general, if there are *O* objects in a scene, and each object has *F* features, then there will be *OF* simple features. There will be *O* conjunctive features that combine two simple features (i.e., *pairwise* conjunctive features). If we limit ourselves to simple and pairwise features to explain the pattern of similarity judgments in Fig. 3, we still will require $OF(F+1)/2$ features per scene, or $OF(F+1)$ features for two scenes that are compared to one another.

Thus, featural approaches to similarity require a fairly large number of features to represent scenes that are organized into parts. Similar problems exist for dimensional accounts of similarity. The situation for these models becomes much worse when we consider that similarity is also influenced by relations between features such as "Black to the left of white" and "square to the left of white." Considering only binary relations, there are $O^2F^2R–OFR$ relations within a scene that contains *O* objects, F features per object, and R different types of relations between features. More sophisticated objections have been raised about these approaches by John Hummel and colleagues (Doumas and Hummel 2005; Hummel 2000, 2001; Hummel and Biederman 1992; Hummel and Holyoak 1997, 2003; Holyoak and Hummel 2000). At the very least, geometric and featural models apparently require an implausibly large number of attributes to account for the similarity relations between structured, multipart scenes.

2.4 Alignment-Based Models

Partly in response to the difficulties that the previous models have in dealing with structured descriptions, a number of researchers have developed alignment-based models of similarity. In these models, comparison is not just matching features, but

determining how elements correspond to, or align with, one another. Matching features are aligned to the extent that they play similar roles within their entities. For example, a car with a green wheel and a truck with a green hood both share the feature *green*, but this matching feature may not increase their similarity much because the car's wheel does not correspond to the truck's hood. Drawing inspiration from work on analogical reasoning (Gentner 1983, Holyoak 2005; Holyoak and Thagard 1995), in alignment-based models, matching features influence similarity more if they belong to parts that are placed in correspondence and parts tend to be placed in correspondence if they have many features in common and are consistent with other emerging correspondences (Goldstone 1994a; Markman and Gentner 1993a). Alignment-based models make purely relational similarity possible (Falkenhainer et al. 1989).

Initial evidence that similarity involves aligning scene descriptions comes from Markman and Gentner's (1993a) result that when subjects are asked to determine corresponding objects, they tend to make more structurally sound choices when they have first judged the similarity of the scenes that contain the objects. Research has found that relational choices such as "smallest object in its set" tend to influence similarity judgments more than absolute attributes like "3 in." when the overall amount of relational coherency across sets is high (Goldstone et al. 1991), the scenes are superficially sparse rather than rich (Gentner and Rattermann 1991; Markman and Gentner 1993a), subjects are given more time to make their judgments (Goldstone and Medin 1994), the judges are adults rather than children (Gentner and Toupin 1986), and abstract relations are initially correlated with concrete relations (Kotovsky and Gentner 1996).

Formal models of alignment-based similarity have been developed to explain how feature matches that belong to well-aligned elements matter more for similarity than matches between poorly aligned elements (Goldstone 1994a; Larkey and Love 2003). Inspired by work in analogical reasoning (Gentner 1983; Holyoak and Thagard 1989), Goldstone's (1994a) SIAM model is a neural network with nodes that represent hypotheses that elements across two scenes correspond to one another. SIAM works by first creating correspondences between the features of scenes. Once features begin to be placed into correspondence, SIAM begins to place objects into correspondence that are consistent with the feature correspondences. Once objects begin to be put in correspondence, activation is fed back down to the feature (mis)matches that are consistent with the object alignments. In this way, object correspondences influence activation of feature correspondences at the same time that feature correspondences influence the activation of object correspondences. Consistent with SIAM (1) aligned-feature matches tend to increase similarity more than unaligned-feature matches (Goldstone 1994a), (2) the differential influence between aligned and unaligned feature matches increases as a function of processing time (Goldstone and Medin 1994), (3) this same differential influence increases with the clarity of the alignments (Goldstone 1994a), and (4) under some circumstances, adding a poorly aligned feature match can actually decrease similarity by interfering with the development of proper alignments (Goldstone 1996). The first effect is shown in Fig. 4. Participants were asked to

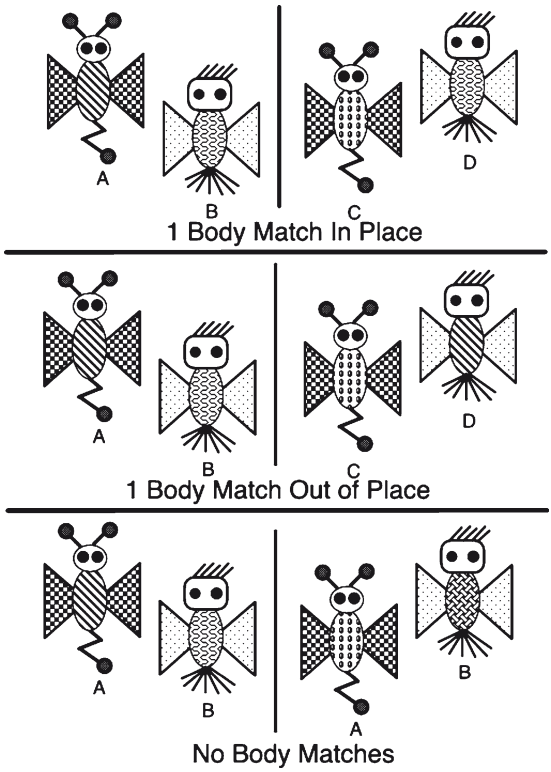

Fig. 4 Sample scenes from Goldstone (1994a). In the *top panel*, the two butterflies that share a matching body pattern are aligned with each other. In the *middle panel*, they are not unaligned. In the *lowest panel*, there are no matching body patterns. Assessments of similarity between scenes decreases as we descend the panels

judge the similarity of scenes made up of two butterflies. The average similarity for the top panel comparison is greater than the middle panel comparison, because the weighting of feature match is affected by its alignment. In the top panel, the matching body pattern occurs between butterflies that are likely to be placed into alignment on the basis of their other feature matches. However, typically the unaligned feature matches (Matches Out of Place) still increase similarity somewhat, and hence the average similarity is higher for the middle than lowest panel comparisons.

Another empirically validated set of predictions stemming from an alignment-based approach to similarity concerns alignable and nonalignable differences (Markman and Gentner 1993b). Nonalignable differences between two entities are attributes of one entity that have no corresponding attribute in the other entity. Alignable differences are differences that require that the elements of the entities first be placed in correspondence. When comparing a police car to an ambulance, a nonalignable difference is that police cars have weapons in them, but ambulances do not. There is no clear equivalent of weapons in the ambulance. Alignable differences include the following: police cars carry criminals to jails rather than carrying sick people to hospitals, a police car is a car while ambulances are vans,

and police car drivers are policemen rather than emergency medical technicians. Consistent with the role of structural alignment in similarity comparisons, alignable differences influence similarity more than nonalignable differences do (Markman and Gentner 1996), and are more likely to be encoded in memory (Markman and Gentner 1997). Alignable differences between objects also play a disproportionately large role in distinguishing between different basic-level categories (e.g., cats and dogs) that belong to the same superordinate category (e.g., animals) (Markman and Wisniewski 1997). In short, knowing these correspondences affects not only how much a matching element increases similarity (Goldstone 1994a), but also how much a mismatching element decreases similarity. Considerable recent research has documented the role of structural alignment in influencing similarity of more natural stimuli, including words (Bernstein et al. 1994; Frisch et al. 1995; Hahn and Bailey 2005), sentences (Bassok and Medin 1997), consumer products (Zhang and Markman 1998), and legal cases (Hahn and Chater 1998; Simon and Holyoak 2002).

2.5 *Transformational Models*

A final historic approach to similarity that has been recently resuscitated is that the comparison process proceeds by transforming one representation into the other. A critical step for these models is to specify what transformational operations are possible.

In an early incarnation of a transformational approach to cognition broadly construed, Garner (1974) stressed the notion of stimuli that are transformationally equivalent and are consequently possible alternatives for each other. In artificial intelligence, Shimon Ullman (1996) has argued that objects are recognized by being aligned with memorized pictorial descriptions. Once an unknown object has been aligned with all candidate models, the best match to the viewed object is selected. The alignment operations rotate, scale, translate, and topographically warp object descriptions.

In transformational accounts that are explicitly designed to model similarity data, similarity is usually defined in terms of transformational distance. In Wiener-Ehrlich et al. (1980) generative representation system, subjects are assumed to possess an elementary set of transformations, and invoke these transformations when analyzing stimuli. Their subjects saw linear pairs of stimuli such as {*ABCD*, *DABC*} or two-dimensional stimuli such as $\left\{\begin{matrix} AB, DA \\ CD \;\; BC \end{matrix}\right\}$. Subjects were required to rate the similarity of the pairs. The researchers determined transformations that accounted for each subjects' ratings from the set {rotate 90°, rotate 180°, rotate 270°, horizontal reflection, vertical reflection, positive diagonal reflection, negative diagonal reflection}. Similarity was assumed to decrease monotonically as the number of transformations required to make one sequence identical to the other increased. Imai (1977) makes a similar claim, empirically finding that as the

number of transformations required to make two strings identical increased, so did the strings' dissimilarity.

Recent work has followed up on Imai's research and has generalized it to stimulus materials including arrangements of Lego bricks, geometric complexes, and sets of colored circles (Hahn et al. 2003). According to these researchers' account, the similarity between two entities is a function of the complexity of the transformation from one to the other. The simpler the transformation, the more similar they are assumed to be. The complexity of a transformation is determined in accord with Kolmogorov complexity theory (Li and Vitanyi 1997), according to which the complexity of a representation is the length of the shortest computer program that can generate that representation. For example, the conditional Kolmogorov complexity between the sequence 1 2 3 4 5 6 7 8 and 2 3 4 5 6 7 8 9 is small, because the simple instructions "add 1 to each digit" and "subtract 1 from each digit" suffice to transform one into the other. Experiments by Hahn et al. demonstrate that once reasonable vocabularies of transformation are postulated, transformational complexity does indeed predict subjective similarity ratings.

3 Conclusions

The study of similarity and comparison is typically justified by the argument that so many theories in cognition depend upon similarity as a theoretical construct. An account of what make problems, memories, objects, and words similar to one another often provides the backbone for our theories of problem solving, attention, perception, and cognition. As William James put it, "This sense of Sameness is the very keel and backbone of our thinking" (James 1890/1950; p. 459).

However, others have argued that similarity is not flexible enough to provide a sufficient account, although it may be a necessary component. There have been many empirical demonstrations of apparent dissociations between similarity and other cognitive processes, most notably categorization. Researchers have argued that cognition is frequently based on theories (Murphy and Medin 1985), rules (Smith and Sloman 1994; Sloman 1996), or strategies that go beyond "mere" similarity (Rips 1989).

Despite the growing body of evidence that similarity comparisons do not always track categorization decisions, there are still some reasons to be sanguine about the continued explanatory relevance of similarity. Categorization itself may not be completely flexible. People are influenced by similarity despite the subjects' intentions and the experimenters' instructions (Allen and Brooks 1991; Palmeri 1997; Smith and Sloman 1994). People seem to have difficulties ignoring similarities between old and new patterns, even when they know a straightforward and perfectly accurate categorization rule. There appears to be a mandatory consideration of similarity in many categorization judgments (Goldstone 1994b).

Similarity and comparison play powerful roles in cognition in situations where we do not know in advance exactly what properties of a situation are critical for its

properties. We rely on comparison to generate inferences and categorize objects into kinds when we do not know exactly what properties are relevant, or when we cannot easily separate an object into separate properties. Accordingly, comparison is an excellent general purpose cognitive strategy. For example, even if we do not know why sparrows have hollow bones, by comparing sparrows to warblers, we may be led to infer that if sparrows have hollow bones, then probably warblers do as well because of their similarity to sparrows. Similarities revealed through comparison thus play a crucial role in making predictions because, tautologically, similar things usually look and behave similarly. Furthermore, once sparrows and warblers are compared, we may not only come to realize that they share the property of hollow bones, but we may even generate an explanation for this trait involving weight, energy requirements to lift a mass, and the importance of flight for the ecological niche of birds. This explanation can cause us to look at birds in a new way. For this reason, comparison not only takes representations as inputs to establish similarities, but also uses similarity to establish new representations (Hofstadter 1997; Medin et al. 1993; Mitchell 1993). When we compare entities, our understanding of the entities changes, and this may turn out to be a far more important consequence of comparison than simply deriving an assessment of similarity.

Author Notes This research was funded by National Science Foundation REESE grant DRL-0910218. Correspondence concerning this chapter should be addressed to rgoldsto@indiana.edu or Robert Goldstone, Psychological and Brain Sciences Department, Indiana University, Bloomington, Indiana 47405. Further information about the laboratory can be found at http://cognitrn.psych.indiana.edu.

References

Allen SW, Brooks LR (1991) Specializing the operation of an explicit rule. J Exp Psychol Gen 120:3–19

Attneave F (1950) Dimensions of similarity. Am J Psychol 63:516–556

Bassok M, Medin DL (1997) Birds of a feather flock together: similarity judgments with semantically rich stimuli. J Mem Lang 36:311–336

Bernstein LE, Demorest ME, Eberhardt SP (1994) A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus/response alignment. J Acoust Soc Am 95:3617–3622

Bransford JD, Schwartz DL (1999) Rethinking transfer: a simple proposal with multiple implications. Rev Res Educ 24:61–100

Burgess C, Lund K (2000) The dynamics of meaning in memory. In: Diettrich E, Markman AB (eds) Cognitive dynamics: conceptual change in humans and machines. Lawrence Erlbaum, Mahwah, NJ, pp 117–156

Bush RR, Mosteller F (1951) A model for stimulus generalization and discrimination. Psychol Rev 58:413–423

Carroll JD, Wish M (1974) Models and methods for three-way multidimensional scaling. In Krantz DH, Atkinson RC, Luce RD, Suppes P (eds) Contemporary developments in mathematical psychology, vol 2. Freeman, San Francisco, pp 57–105

Collins AM, Quillian MR (1969) Retrieval time from semantic memory. J Verbal Learn Verbal Behav 8:240–247

Doumas LAA, Hummel JE (2005) Approaches to modeling human mental representation: what works, what doesn't, and why. In Holyoak KJ, Morrison RG (eds) The Cambridge handbook of thinking and reasoning. Cambridge University Press, Cambridge, England, pp 73–91
Eisler H, Ekman G (1959) A mechanism of subjective similarity. Acta Psychol 16:1–10
Falkenhainer B, Forbus KD, Gentner D (1989) The structure-mapping engine: Algorithm and examples. Artif Intell 41:1–63
Frisch SA, Broe MB, Pierrehumbert JB (1995) The role of similarity in phonology: Explaining OCP-Place. In Elenius K, Branderud P (eds) Proceedings of the, 13th International Conference of the Phonetic Sciences, Stockholm, vol 3, pp 544–547
Garner WR (1974) The processing of information and structure. Wiley, New York
Gati I, Tversky A (1984) Weighting common and distinctive features in perceptual and conceptual judgments. Cogn Psychol 16:341–370
Gentner D (1983) Structure-mapping: a theoretical framework for analogy. Cogn Sci 7:155–170
Gentner D, Namy L (1999) Comparison in the development of categories. Cogn Dev 14:487–513
Gentner D, Rattermann MJ (1991) Language and the career of similarity. In: Gelman SA, Byrnes JP (eds) Perspectives on language and thought interrelations in development. Cambridge University Press, Cambridge, England
Gentner D, Toupin C (1986) Systematicity and surface similarity in the development of analogy. Cogn Sci 10(3):277–300
Gentner D, Loewenstein J, Thompson L (2003) Learning and transfer: a general role for analogical encoding. J Educ Psychol 95:393–408
Gick ML, Holyoak KJ (1983) Schema induction and analogical transfer. Cogn Psychol 15:1–38
Gluck MA (1991) Stimulus generalization and representation in adaptive network models of category learning. Psychol Sci 2:50–55
Goldstone RL (1994a) Similarity, interactive activation, and mapping. J Exp Psychol Learn Mem Cogn 20:3–28
Goldstone RL (1994b) The role of similarity in categorization: Providing a groundwork. Cognition 52:125–157
Goldstone RL (1996) Alignment-based nonmonotonicities in similarity. J Exp Psychol Learn Mem Cogn 22:988–1001
Goldstone RL, Medin DL (1994) The time course of comparison. J Exp Psychol Learn Mem Cogn 20:29–50
Goldstone RL, Medin DL, Gentner D (1991) Relations, attributes, and the non-independence of features in similarity judgments. Cogn Psychol 23:222–264
Goldstone RL, Medin DL, Halberstadt J (1997) Similarity in context. Mem Cogn 25:237–255
Griffiths TL, Steyvers M, Tenenbaum JBT (2007) Topics in semantic representation. Psychol Rev 114(2):211–244
Hahn U (2003) Similarity. In: Nadel L (ed) Encyclopedia of cognitive science. Macmillan, London
Hahn U, Bailey RM (2005) What makes words sound similar? Cognition 97:227–267
Hahn U, Chater N (1998) Understanding similarity: a joint project for psychology, case-based reasoning and law. Artif Intell Rev 12:393–427
Hahn U, Chater N, Richardson LB (2003) Similarity as transformation. Cognition 87:1–32
Hayes-Roth B, Hayes-Roth F (1977) Concept learning and the recognition and classification of exemplars. J Verbal Learn Verbal Behav 16:321–338
Hintzman DL (1986) Schema abstraction in a multiple-trace memory model. Psychol Rev 93:411–428
Hofstadter D (1997) Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought. Basic Books, New York
Holyoak KJ (2005) Analogy. In: Holyoak KJ, Morrison RG (eds) The Cambridge Handbook of Thinking and Reasoning. Cambridge University Press, Cambridge, UK

Holyoak KJ, Hummel JE (2000) The proper treatment of symbols in a connectionist architecture. In: Dietrich E, Markman A (eds) Cognitive dynamics: conceptual change in humans and machines. Erlbaum, Hillsdale, NJ

Holyoak KJ, Koh K (1987) Surface and structural similarity in analogical transfer. Mem Cogn 15:332–340

Holyoak KJ, Thagard P (1989) Analogical mapping by constraint satisfaction. Cogn Sci 13:295–355

Holyoak KJ, Thagard P (1995) Mental leaps: analogy in creative thought. MIT, Cambridge, MA

Horgan DD, Millis K, Neimeyer RA (1989) Cognitive reorganization and the development of chesss expertise. Int J Pers Construct Psychol 2:15–36

Hubel DH, Wiesel TN (1968). Receptive fields and functional architecture of monkey striate cortex. J Physiol 195:215–243

Huber J, Payne JW, Puto C (1982) Adding asymmetrically dominated alternatives: violations of regularity and the similarity hypothesis. J Consum Res 9:90–98

Hummel JE (2000) Where view-based theories break down: the role of structure in shape perception and object recognition. In: Dietrich E, Markman A (eds) Cognitive dynamics: conceptual change in humans and machines. Erlbaum, Hillsdale, NJ

Hummel JE (2001) Complementary solutions to the binding problem in vision: implications for shape perception and object recognition. Vis Cogn 8:489–517

Hummel JE, Biederman I (1992) Dynamic binding in a neural network for shape recognition. Psychol Rev 99:480–517

Hummel JE, Holyoak KJ (1997) Distributed representations of structure: a theory of analogical access and mapping. Psychol Rev 104:427–466

Hummel JE, Holyoak KJ (2003) A symbolic-connectionist theory of relational inference and generalization. Psychol Rev 110:220–263

Imai S (1977) Pattern similarity and cognitive transformations. Acta Psychol 41:433–447

Jakobson R, Fant G, Halle M (1963) Preliminaries to speech analysis : the distinctive features and their correlates. MIT, Cambridge, MA

James W (1890/1950) The principles of psychology. Dover, New York (Original work published 1890)

Katz JJ, Fodor J (1963) The structure of semantic theory. Language 39:170–210

Kotovsky L, Gentner D (1996) Comparison and categorization in the development of relational similarity. Child Dev 67:2797–2822

Landauer TK, Dumais ST (1997) A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychol Rev 104:211–240

Larkey LB, Love BC (2003) CAB: connectionist analogy builder. Cogn Sci 27:781–794

Lee MD (2002) A simple method for generating additive clustering models with limited complexity. Mach Learn 49:39–58

Li M, Vitanyi P (1997) An introduction to Kolmogorov complexity and its applications, 2nd edn. Springer, New York

Markman AB, Gentner D (1993a) Structural alignment during similarity comparisons. Cogn Psychol 25:431–467

Markman AB, Gentner D (1993b) Splitting the differences: a structural alignment view of similarity. J Mem Lang 32:517–535

Markman AB, Gentner D (1996) Commonalities and differences in similarity comparisons. Mem Cogn 24:235–249

Markman AB, Gentner D (1997) The effects of alignability on memory. Psychol Sci 8:363–367

Markman AB, Wisniewski EJ (1997) Similar and different: the differentiation of basic-level categories. J Exp Psychol Learn Mem Cogn 23:54–70

Medin DL, Shaffer MM (1978) A context theory of classification learning. Psychol Rev 85:207–238

Medin DL, Goldstone RL, Gentner D (1993) Respects for similarity. Psychol Rev 100:254–278
Mitchell M (1993) Analogy-making as perception: a computer model. MIT, Cambridge, MA
Murphy GL, Medin DL (1985) The role of theories in conceptual coherence. Psychol Rev 92:289–316
Namy LL, Gentner D (2002) Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. J Exp Psychol Gen 131:5–15
Navarro DJ, Griffiths TL (2007) A nonparametric Bayesian method for inferring features from similarity judgments. Adv Neural Inform Process Syst 19:1033–1040
Navarro DJ, Lee MD (2004) Common and distinctive features in stimulus representation: A modified version of the contrast model. Psychon Bull Rev 11(6):961–974
Neisser U (1967) Cognitive psychology. Appleton-Century-Crofts, New York
Nosofsky RM (1984) Choice, similarity, and the context theory of classification. J Exp Psychol Learn Mem Cogn 10:104–114
Osherson D, Smith EE, Wilkie O, Lopez A, Shafir E (1990) Category-based induction. Psychol Rev 97:185–200
Osterholm K, Woods DJ, Le Unes A (1985) Multidimensional scaling of Rorschach inkblots: Relationships with structured self-report. Pers Individ Dif 6:77–82
Palmer SE (1975) Visual perception and world knowledge. In: Norman DA, Rumelhart DE (eds) Explorations in cognition. Freeman, San Francisco
Palmeri TJ (1997) Exemplar similarity and the development of automaticity. J Exp Psychol Learn Mem Cogn 23:324–354
Polk TA, Behensky C, Gonzalez R, Smith EE (2002) Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency. Cognition 82:B75–B88
Reed SK (1972) Pattern recognition and categorization. Cogn Psychol 3:382–407
Rips LJ (1989) Similarity, typicality, and categorization. In: Vosniadu S, Ortony A (eds) Similarity, analogy, and thought. Cambridge University Press, Cambridge, pp 21–59
Rips LJ, Shoben EJ, Smith EE (1973) Semantic distance and the verification of semantic relationships. Journal of Verbal Learning and Verbal Behavior, 12:1–20
Ritov I, Gati I, Tversky A (1990) Differential weighting of common and distinctive components. J Exp Psychol Gen 119:30
Ross BH (1989) Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. J Exp Psychol Learn Mem Cogn 15:456–468
Schvaneveldt RW, Durso FT, Goldsmith TE, Breen TJ, Cooke NM, Tucker RG, DeMaio JC (1985) Measuring the structure of expertise. Int J Man-Mach Stud 23:699–728
Shanon B (1988) On similarity of features. New Ideas Psychol 6:307–321
Shepard RN (1962a) The analysis of proximities: multidimensional scaling with an unknown distance function. Part I. Psychometrika 27:125–140
Shepard RN (1962b) The analysis of proximities: multidimensional scaling with an unknown distance function. Part II. Psychometrika 27:219–246
Shepard RN (1972) Psychological representation of speech sounds. In: David EE Jr, Denes PB (eds) Human communication: a unified view. McGraw-Hill, New York
Shepard RN (1982) Geometrical approximations to the structure of musical pitch. Psychol Rev 89:305–333
Shepard RN (1987) Toward a universal law of generalization for psychological science. Science 237:1317–1323
Shepard RN, Arabie P (1979) Additive clustering: representation of similarities as combinations of discrete overlapping properties. Psychol Rev 86:87–123
Shiffrin RM, Steyvers M (1997) A model for recognition memory: REM: retrieving effectively from memory. Psychon Bull Rev 4(2):145–166
Simon D, Holyoak KJ (2002) Structural dynamics of cognition: From consistency theories to constraint satisfaction. Pers Soc Psychol Rev 6:283–294
Simonson I (1989) Choice based on reasons: the case of attraction and compromise effects. J Consum Res 16:158–174

Sjoberg L (1972) A cognitive theory of similarity. Goteborg Psychol Rep 2(10)
Sloman SA (1996) The empirical case for two systems of reasoning. Psychol Bull 119:3–22
Smith EE, Sloman SA (1994) Similarity-versus rule-based categorization. Mem Cogn 22:377–386
Smith EE, Shoben EJ, Rips LJ (1974) Structure and process in semantic memory: a featural model for semantic decisions. Psychol Rev 81:214–241
Tenenbaum JB (1996) Learning the structure of similarity. In: Tesauro G, Touretzky DS, Leen TK (eds) Advances in neural information processing systems, 8. MIT, Cambridge, MA, pp 4–9
Tenenbaum JB, De Silva V, Lanford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:22–23
Torgerson WS (1965) Multidimensionsal scaling of similarity. Psychometrika 30:379–393
Treisman AM (1986) Features and objects in visual processing. Sci Am 255:106–115
Tversky A (1977) Features of similarity. Psychol Rev 84:327–352
Tversky A, Gati I (1982) Similarity, separability, and the triangle inequality. Psychol Rev 89:123–154
Tversky A, Hutchinson JW (1986) Nearest neighbor analysis of psychological spaces. Psychol Rev 93:3–22
Ullman S (1996) High-level vision: object recognition and visual cognition. MIT, London
Wiener-Ehrlich WK, Bart WM, Millward R (1980) An analysis of generative representation systems. J Math Psychol 21(3):219–246
Winston PH (1975) Learning structural descriptions from examples. In: Winston PH (ed) The psychology of computer vision. McGraw-Hill, New York
Zhang S, Markman AB (1998) Overcoming the early entrant advantage: the role of alignable and nonalignable differences. J Market Res 35:413–426

Causal Thinking

Michael R. Waldmann

Abstract The ability to acquire and reason with causal knowledge belongs to our most central cognitive competencies. Causal knowledge serves various functions: It enables us to predict future events, to diagnose the causes of observed events, and to choose the right actions to achieve our goals. The chapter gives an overview of the causal-model approach to causal reasoning and learning. It focuses on the contrast between traditional associationist theories and this more recent rational approach to causal thinking, and discusses this theory in light of recent experimental evidence.

1 Introduction

The ability of people to predict future events, to explain past events, and to choose appropriate actions to achieve goals belongs to the most central cognitive competencies that allow us to be successful agents in the world. How is knowledge about regularities in the world learned, stored, and accessed? A plausible theory that governs our intuitive thinking assumes that *causality* is the "cement of the universe" (Mackie 1974), which underlies the orderly relations between events. According to this view some event types, causes, have the capacity or power to generate their effects through hidden mechanisms.

The philosopher David Hume questioned this view in his seminal writings (e.g., Hume 1748/1977). He looked at situations in which he observed causal relations, and did not detect any empirical features that might correspond to evidence for causal powers. What he found instead was spatio-temporally ordered successions of event pairs, but nothing beyond that might correspond to power.

The psychology of learning has adopted Hume's view by focusing on spatio-temporally ordered events. According to many learning theories, causal predictions are driven by associative relations that have been learned on the basis

M.R. Waldmann
Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073, Göttingen, Germany
e-mail: michael.waldmann@bio.uni-goettingen.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_8,

of observed covariations between events (e.g., Allan 1993; Shanks and Dickinson 1987). Similar to Pavlov's dog which has learned to predict food when it hears a tone (i.e., classical conditioning), or to a pigeon's learning that a lever press produces food (i.e., instrumental conditioning), we learn about predictive relations between observed events or interventional relations between our acts and their outcomes. According to the associationist view, there is no need for the concept of causality. Thus, consistent with the epistemology of empiricism the concept of causality as referring to causal powers was dropped and replaced by predictive covariational relations between observable events.

Hume's analysis leaves us with a puzzle. On the one hand, he seems to have correctly observed that covariations between observable events are the primary perceptual input for causal inductions. On the other hand, many researchers agree that when thinking about the world we go beyond the information given by assuming hidden capacities, forces, or processes beyond the surface of orderly event successions (Ahn et al. 1995; Cheng 1997). The following overview will summarize research showing that people (and some animals) indeed go beyond covariations in causal learning and reasoning.

2 Causal-Model Theory: Beyond Covariations

Hume seems to have correctly observed that the *input of causal learning* largely consists of covariation information. It can be shown, however, that mental representations that merely mirror this input cannot explain the competencies people have when dealing with causal situations (see also Buehner and Cheng 2005; Waldmann et al. 2006; Waldmann et al. 2008; Waldmann 1996, for overviews of causal-model theory). If we had no *causal* knowledge we could not represent the difference between causal and noncausal spurious statistical relations. For example, in the common-cause model in Fig. 1 the *Midosis* virus is a direct cause of two effects, symptom 1 and symptom 2. Inserting this virus into an organism would produce both effects. However, causing symptom 1 by other means would not affect symptom 2. Whereas the virus is causally related to either effect, the two effects only covary but are not causally related (i.e., spurious noncausal correlation).

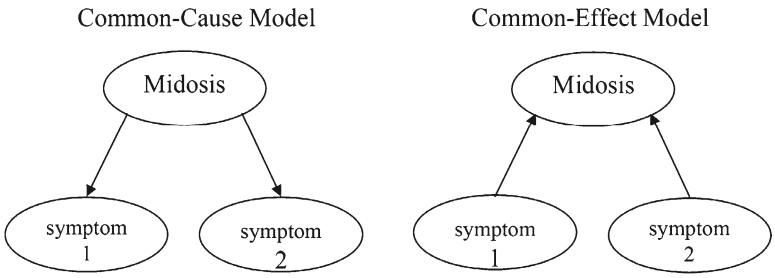

Fig. 1 Examples of a common-cause and common-effect model. The *arrows* represent causal relations directed from causes to effects

The difference between these two types of relations cannot be represented by theories that are only sensitive to statistical covariations (e.g., associative theories), whereas they are of utmost importance when we reason about the outcomes of interventions.

Covariational knowledge also fails to make the fundamental distinction between causes and effects, which is central for causality. Whereas causes are typically correlated with effects, and effects are typically correlated with causes, causal relations are asymmetric. Causes generate effects and not vice versa. Again, this distinction is important when we reason about causal systems. We can use information about causes to predict effects, and information about effects to diagnose their causes. However, we can only intervene in causes to achieve effects but not vice versa. Again we need to go beyond covariational information to correctly represent this knowledge.

Finally, causal models entail statistical relations between events that are helpful in learning. For example, multiple effects of a common cause but not multiple causes of a common effect are correlated in a predictable way. These correlations entailed by the structure of causal models allow us to limit the complexity of the representations we need to represent the covariations entailed by causal knowledge.

How can Hume's insight that the observational input only offers covariation information be reconciled with the observation that we represent our world in terms of causal models endowed with powers and mechanisms? A possible answer is that Hume's empiricist epistemology was incomplete. As many philosophers of science have revealed, apart from concepts referring to observable events (e.g., covariations) our theories also contain theoretical concepts which are only indirectly tied to the observable data (e.g., causal powers). Theoretical concepts are components of theories which specify how they can be estimated from data in a specific situation. Causal-model theory and related accounts claim that we have a tendency to assume the existence of deep causal relations behind the visible surface, which leads us to align covariational input with causal-model representations (see also Buehner and Cheng 2005). Similar to using sparse visual input to induce 3D object representations that go beyond the learning input, people have a tendency to represent some events as causes with the power to generate or prevent effects, and they build causal networks that can be used for inferences and planning (see Cheng 1997; Tenenbaum and Griffiths 2003; Waldmann et al. 2006, 2008).

In the following sections, I will first discuss empirical studies testing causal-model theory showing that humans (and some nonhuman animals) indeed go beyond covariations to infer causal structure (see also Gopnik et al. 2004; Lu et al. 2008; Sloman 2005; Tenenbaum and Griffiths 2003, for related views). In the final section I will add some speculations on the processes underlying these competencies.

2.1 Sensitivity to the Asymmetry of Causes and Effects

The distinction between causes and effects is a central feature of causal representations. Thus, one line of our research focused on whether there is evidence that people are sensitive to this distinction (see Waldmann 1996, for a summary of early work).

Fenker et al. (2005) have investigated this question in a *semantic memory task*. We were interested in whether causal relations are represented and accessed differently from associative relations in semantic memory. In the experiments, participants were shown pairs of words, one after another, either describing events that referred to a cause (e.g., spark) or an effect (e.g., fire) of a causal relation. Both the temporal order of word presentation and the question to which participants had to respond was manipulated. Interestingly, when we asked whether the two events are *causally* related, participants answered faster when the first word referred to a cause and the second word to its effect than vice versa. No such asymmetry was observed, however, when we asked about associative relations. The associative questions probed participants whether the words describing the events are related in some meaningful way. People appear to distinguish the roles of cause and effect when queried specifically about a causal relation, but not when the same information is evaluated for the presence of an associative relation.

In a follow-up study, a functional magnetic resonance imaging experiment was performed to investigate the hypothesized dissociation between the use of semantic knowledge to evaluate specifically causal relations in contrast to general associative relations (Satpute et al. 2005). Again, identical pairs of words were judged for causal or associative relations in different blocks of trials. Causal judgments, beyond associative judgments, generated distinct activation in left dorsolateral prefrontal cortex and right precuneus. These findings indicate that the evaluation of causal relations in semantic memory involves additional neural mechanisms relative to those required to evaluate associative relations.

Whereas semantic memory tasks target the results of learning, there is also evidence that people go beyond covariations in trial-by-trial learning tasks (see Waldmann and Holyoak 1992; Waldmann et al. 1995; Waldmann 2000). The general paradigm presents participants in different conditions with identical covarying events. If Hume was right, and learning simply consists of processing these observed spatio-temporally ordered covariations, the outcome of the learning process should be the same. However, if participants go beyond covariations and form representations of the underlying causal models, their reasoning should be sensitive to the structure of these models. In one study Waldmann (2001) presented learners first with cues that represented substances in hypothetical patients' blood and then gave feedback about fictitious hematological diseases (e.g., *Midosis*). Two conditions manipulated – through initial instructions – whether learners interpreted the substances (i.e., cues) as effects of the diseases (common-cause model) or as causes (common-effect model)(see Fig. 1). In the common-cause condition, participants were told that the diseases caused some of the substances in the blood which could be used to diagnose the diseases. In contrast, in the common-effect condition the very same substances were described as coming from food items, which were suspected to be causing the novel blood diseases.

According to associative learning theories, learners should learn to predict the diseases from information about the presence or absence of the substances, which was always provided first as cues in the learning trials. Since cues and outcomes were identical in both conditions, the learning process should be identical. In contrast, causal-model theory predicts that learners are sensitive to the distinction between

cues that represent causes versus effects, which should influence the learning and reasoning processes. The results showed that causal models indeed affected reasoning. Learners treated the substances as potentially competing explanations of the disease in the common-effect condition, whereas the substances were treated as collateral, collaborating effects of a common cause in the contrasting condition. Thus, despite the fact that all learners observed the same sequence of events, they assigned different causal roles to these events, and consequently made different inferences.

2.2 *Predicting Outcomes of Hypothetical Observations Versus Hypothetical Interventions*

Predicting or diagnosing on the basis of real or hypothetical observed events (e.g., observed substances) are both examples of observational inferences. We may also be confronted with the task to predict or diagnose on the basis of hypothetical interventions. Sometimes these two types of predictions coincide, but very often they do not. For example, hypothetically observing symptom 1 in the common-cause model depicted in Fig. 1 allows us to diagnose the *Midosis* virus and infer symptom 2 from there. However, a hypothetical intervention that causes symptom 1 by other means should not change our inferences about the virus and the other symptom. Again, to make correct inferences here, the learner needs to go beyond the given covariations, and assign causal roles to the observed events.

An associative learning theorist might react to this example arguing that human and nonhuman animals could distinguish between observing and intervening on the basis of observational (i.e., classical) and instrumental conditioning. We may, for example, have learned that symptom 1 predicts symptom 2 in an observational learning setting and in parallel may have tried to cause symptom 1 by other means with no effect on the virus and the other symptom. This solution only works, however, if learners are provided with both kinds of learning experiences, not if they only passively observe covarying events and then are requested to make both observational and interventional predictions.

Waldmann and Hagmayer (2005) tested people's competence to derive predictions for hypothetical observations and hypothetical interventions from causal models that had been learned purely through observation (see also Hagmayer et al. 2007; Sloman and Lagnado 2005). In a fictitious scenario, participants were either told that scientists hypothesized that the three hormones *pixin*, *sonin*, and *xanthan* form a common-cause or a causal chain model in animals (see Fig. 2). All participants in the two conditions received identical observational data indicating that the three hormones were connected by probabilistic causal relations. In the subsequent test phase learners were asked to make predictions about hypothetical observations of *sonin* in new animals, and hypothetical interventions, which increased *sonin* in the blood of new animals by means of inoculations. (In other conditions the *sonin* levels were hypothetically decreased.)

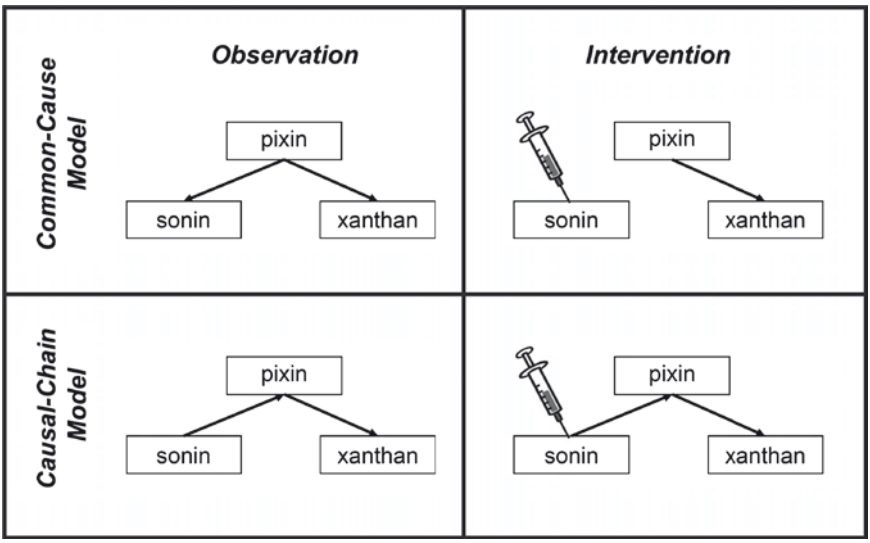

Fig. 2 Observational and interventional predictions in a common-cause and causal-chain model, in which three hormones are causally connected. The *left side* shows the models presented in the learning phase which can be used for observational predictions. The *right side* depicts the models underlying the predictions of the outcomes of hypothetical interventions. An intervention in *sonin* leads to full discounting of *pixin* in the common-cause but not in the chain model. Full discounting can be expressed by removing the *arrow* from *pixin* to *sonin*, which turns these two substances statistically independent in the test situation

The observational inferences can be modeled on the basis of the two presented causal models. Since the three events are statistically related in both causal models, the observation of the presence of *sonin* allows us to reason that *pixin* and consequently *xanthan* are also very likely to be present. Interventional predictions entail different predictions in one of the models. An intervention that adds *sonin* to the blood leads in the common-cause model to the consequence that the levels of *sonin* are now determined by this intervention and no longer by its usual cause (*pixin*), whose causal influence is preempted by the novel intervention (see Spirtes et al. 1993; Pearl 2000; Woodward 2003). One way to model this intervention is to remove the arrow from *sonin's* normal cause, *pixin*, that is being explained away by the new intervention (see also Waldmann et al. 2008, for a more general theory). The removal expresses that *pixin* is no longer a cause of *sonin* (see Fig. 2, right) in the test situation. Due to the removal of the arrow in the common-cause model *sonin* becomes independent of *xanthan* so that regardless of whether *sonin* is increased or decreased by an intervention, the level of *xanthan* should remain at an identical level.

The chain condition served as a control that showed that observing and intervening do not always lead to different predictions (see Fig. 2). Since there are no alternative causes of *sonin* that are being discounted, there should be no difference whether *sonin* is hypothetically observed or generated by an intervention in this

model. As a consequence, participants should make identical predictions for the hypothetical observational and interventional questions in the chain condition. In our experiments, participants' responses corresponded to these predictions remarkably well. They were capable of predicting patterns they had never observed, which indicates that despite identical learning input they used causal model representations to transform identical covariational information into different types of predictions. In several further experiments we manipulated the statistical parameters of the models (base rates of the events; causal strength of causal links) and showed that participants' predictions were not only driven by the structure of the causal models but also by the learned parameters (see also Meder et al. 2008, 2009, for related evidence in trial-by-trial learning tasks).

2.2.1 Causal Reasoning in Nonhuman Animals

Humans apparently have the natural capacity to form causal representations. How about nonhuman animals? A number of researchers have asserted that causal reasoning and learning are faculties that form a dividing line between humans and nonhuman animals (Povinelli 2000; Tomasello and Call 1997). Recent research by Blaisdell et al. (2006) casts doubt on this conclusion (see also Beckers et al. 2006).

In one experiment, rats went through a purely observational learning phase in which the light was sometimes followed by a tone and at other times followed by food. Importantly, no instrumental learning took place. When in the subsequent observational test phase the rats again heard the tone, they showed that they expected food in the niche in which it was typically delivered. Apparently they reasoned through the causal model link-by-link from the tone through the light to the probable presence of food (see also Waldmann et al. 2008). In contrast, in a second test a lever which the rats had never seen before was introduced into the cage. Whenever the rats curiously pressed the lever, the same tone was presented. Now, although tone and food had been associated by the rats in the learning phase as indicated in the observational test phase, the rats were less inclined to search for food after the lever presses. Apparently they reasoned that they – and not the light – were the cause of the tone, which led to their reluctance to expect food.

In a second study of Blaisdell et al. (2006) a causal chain was presented in which the tone preceded light which in turn preceded food. Consistent with causal-model theory, the rats expected food regardless of whether they observed the tone or generated it with the lever. This shows that they were not generally reluctant to expect food after a novel intervention. The results revealed an understanding of causal relations and demonstrate that rats correctly differentiated between observing and intervening and different causal models.

Whereas associative theories predict associations between tone and food regardless of whether the tone is observed or generated by an action in the test phase, causal-model theory predicts that the intervention at test should be viewed as a potent alternative cause of the tone. Leising et al. (2008) report further tests of causal-model theory. One key prediction is that full discounting of the light should only be

observed when the new alternative cause is viewed as deterministic and independent of the previous cause, the light. Independence and determinism are two hallmark features of interventions but not necessarily of arbitrary events. Consequently, we (Leising et al. 2008) only observed discounting with interventions but not with other observable events. Moreover, rats were capable of flexibly switching between observational and interventional predictions. These results confirm that rats are capable of flexible causal reasoning.

Although this research documents remarkable causal competencies in rats, it nevertheless leaves some interesting questions open. It is true that rats in the two sets of experiments were capable of correctly inferring the outcomes of observations and interventions, but they did not display this knowledge in their actions. For example, although the rats strongly expected food when their intervention caused a tone that was directly causing food, they did not increase the number of lever presses to get more food in this situation. It may well be that rats only have partial incomplete knowledge of causal relations that serves their predictive competencies but falls short of underwriting the action system (see also Penn et al. 2008, for a skeptical view).

2.3 Estimating Causal Parameters

The primary difference between different types of causal models, such as common-cause or common-effect models, lies in the way directed causal arrows are combined. The previous sections have discussed studies showing that people are sensitive to the structural aspects of causal models and capable of coordinating identical learning input with different causal structures. Causal models do not only have a structure, the individual links also have attached strengths as parameters that need to be learned. According to associative theories strength corresponds to observed covariations. However, Cheng (1997) has shown that covariations do not directly mirror causal strength (or causal power in her terminology). According to Cheng, causal power is an unobservable property of causes which expresses the probability of causes generating or preventing effects in ideal circumstances in which no other causes are simultaneously present. This information is not directly provided by the learning input because typically multiple events co-occur. However, causal strength can be inferred. Again, learners need to go beyond covariations. Cheng has derived formulas that allow us to infer causal power under some background assumptions.

An example may illustrate why causal power and covariations do not necessarily correspond. Imagine a drug that generates itching as a side effect. If this drug is solely given to patients who already suffer from itching, no covariation would be observed. The probability of itching remains the same regardless of whether the drug is taken or not. However, it is intuitively clear that this situation does not allow us to assess causal power. The drug may still be a strong cause of itching although in this situation it does not display its power. Cheng and colleagues have shown that people take such properties of the learning situation into account when assessing causal power, and hence go beyond covariations (see Buehner et al. 2003).

Causal structures and their parameters are not independent entities but are deeply intertwined. The causal strength between a cause and an effect, for example, needs to be estimated differently depending on whether or not there is a confounding alternative cause. For example, if we learn the causal strength between the virus and a symptom, we need to control for possible confounds, but not for further effects of the symptom. Waldmann and Hagmayer (2001) have shown that learners are indeed sensitive to the causal roles of events when estimating causal strength.

Causal strength of individual links is not the only parameter of causal models that needs to be inferred, there are also different ways in which multiple causes can combine when they jointly cause a common effect. A typical assumption is that the combination of two generative causes increases the likelihood of the effect beyond what either cause would do. This so called "noisy-or" rule is a default assumption of many networks (Cheng 1997; Griffiths and Tenenbaum 2005), which is assumed to hold unless there are reasons to assume that the causes interact (see Novick and Cheng 2004). Waldmann (2007) has studied continuously varying effects, and has shown that people use background domain knowledge when choosing an integration rule. For example, in one experiment causes were differently colored liquids which could cause the increase of the heart rate of animals. When the liquids were described as affecting intensive quantities (e.g., taste) or preferences (e.g., liking) people were biased towards averaging the causal influences, whereas extensive quantities (e.g., liquids represented drugs with different strengths) led to a tendency to add.

2.4 Limitations of Causal Reasoning

Although people exhibit a sophisticated ability to reason with causal models, there is also evidence for limitations. For example, Waldmann and Walker (2005) have shown that people have difficulties with transforming covariation information into causal-model representations when the task is complex, presented abstractly, or when the learner operates at her information processing limit (see also De Houwer and Beckers 2003). Reips and Waldmann (2008) have similarly found that base rates may be neglected in complex learning tasks. Their results showed that learners are capable of incorporating base rate information in their judgments regardless of the direction in which the causal structure is learned. However, this only holds true for relatively simple scenarios. When complexity was increased, base rates were only used after diagnostic learning, but were largely neglected after predictive learning.

3 Inducing Causal Structures

The previous sections have focused on evidence showing that people and some nonhuman animals go beyond covariations to build causal model representations instead of mirroring statistical relations between cues and outcomes in the learning

input. This research demonstrates that we are not tied to the surface of covarying events. However, I did not address the question how people learn to separate causal model representation from statistical learning input. Why do we not just stick to the surface level? So far there is little research addressing this question. Different factors may be at play here. Infants may be born with a natural tendency to interpret causal events as caused by hidden forces, as suggested by Leslie and colleagues (e.g., Leslie and Keeble 1987). Others have suggested that the tendency to interpret events causally may be triggered by infants' experience of their own actions changing events in their environment, which might provide the basis for further causal knowledge (Dickinson and Balleine 2000; White 2006). Most likely both factors are at play, but we do know little about their relative contributions.

Independent of whether our bias to attribute a causal texture to the world is innate or learned, we need to learn to coordinate the learning input with hypothetical causal models. Where do these models come from? Lagnado et al. (2007) have suggested that we use several cues to form hypothetical models which in turn guide the processing of the learning input. The primary role of the learning input is to provide information about the existence and strength of the causal links (i.e., parameter estimation). The cues underlying structure inductions include temporal order (causes typically precede effects), interventions (interventions target causes, not effects), or coherence with prior knowledge or verbal instructions. Often these cues signal the same structure, but occasionally they may be in conflict. For example, a physician may see a symptom (i.e., an effect) prior to testing for its cause, which requires him to disentangle learning from causal order. Lagnado et al. (2007) summarize various experiments exploring how people coordinate different cues to induce causal models.

4 Conclusion

Hume has presented us with a puzzle: How do we acquire causal knowledge when we only observe covariation information? We have reported a number of studies showing that both human and nonhuman animals have a natural tendency to coordinate covariations with deep causal model representations.

One important question for future research is to explore the generality and differences of causal reasoning capacities across species. Another interesting question will be to analyze the relation between causal reasoning and rational models, such as causal Bayes nets (Gopnik et al. 2004; Lu et al. 2008; Griffiths and Tenenbaum 2005). Our findings on limitations of causal reasoning suggest that such models, if interpreted as psychological theories, may often exaggerate what human and nonhuman animals can do (see Waldmann et al. 2008). Answers to these questions promise to elucidate the structure, origin, and evolution of causal reasoning as an invaluable cognitive tool for surviving and succeeding in one's world.

References

Ahn W-K, Kalish CW, Medin DL, Gelman SA (1995) The role of covariation versus mechanism information in causal attribution. Cognition 54:299–352

Allan LG (1993) Human contingency judgment: rule based or associative? Psychol Bull 114:435–448

Beckers T, Miller RR, De Houwer J, Urushihara K (2006) Reasoning rats: forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. J Exp Psychol Gen 135:92–102

Blaisdell AP, Sawa K, Leising KJ, Waldmann MR (2006) Causal reasoning in rats. Science 311:1020–1022

Buehner M, Cheng P (2005) Causal learning. In: Holyoak J, Morrison B (eds) K. Cambridge University Press, The Cambridge handbook of thinking and reasoning. Cambridge, pp 143–168

Buehner MJ, Cheng PW, Clifford D (2003) From covariation to causation: a test of the assumption of causal power. J Exp Psychol Learn Mem Cogn 29:1119–1140

Cheng PW (1997) From covariation to causation: a causal power theory. Psychol Rev 104:367–405

De Houwer J, Beckers T (2003) Secondary task difficulty modulates forward blocking in human contingency learning. Q J Exp Psychol 56B:345–357

Dickinson A, Balleine BW (2000) Causal cognition and goal–directed action. In: Heyes C, Huber L (eds) The evolution of cognition. MIT, Cambridge, MA, pp 185–204

Fenker DB, Waldmann MR, Holyoak KJ (2005) Accessing causal relations in semantic memory. Mem Cogn 33:1036–1046

Gopnik A, Glymour C, Sobel DM, Schulz LE, Kushnir T, Danks D (2004) A theory of causal learning in children: causal maps and Bayes nets. Psychol Rev 111:3–32

Griffiths TL, Tenenbaum JB (2005) Structure and strength in causal induction. Cogn Psychol 51:285–386

Hagmayer Y, Sloman SA, Lagnado DA, Waldmann MR (2007) Causal reasoning through intervention. In Gopnik A, Schultz LE (eds) Causal learning: psychology, philosophy & computation. Oxford University Press, Oxford, pp 86–100

Hume D (1748/1977) An enquiry concerning human understanding. Hackett, Indianapolis

Lagnado DA, Waldmann MR, Hagmayer Y, Sloman SA (2007) Beyond covariation. Cues to causal structure. In: Gopnik A, Schulz L (eds) Causal learning: psychology, philosophy & computation. Oxford University Press, Oxford

Leising KJ, Wong J, Waldmann MR, Blaisdell AP (2008) The special status of actions in causal reasoning in rats. J Exp Psychol Gen 127:514–527

Leslie AM, Keeble S (1987) Do six-month-old infants perceive causality. Cognition 25:265–288

Lu H, Yuille A, Liljeholm M, Cheng PW, Holyoak KJ (2008) Bayesian generic priors for causal learning. Psychol Rev 115:955–984

Mackie JL (1974) The cement of the universe. A study of causation. Clarendon, Oxford

Meder B, Hagmayer Y, Waldmann MR (2008) Inferring interventional predictions from observational learning data. Psychon Bull Rev 15:75–80

Meder B, Hagmayer Y, Waldmann MR (2009) The role of learning data in causal reasoning about observations and interventions. Mem Cogn 37:249–264

Novick LR, Cheng PW (2004) Assessing interactive causal power. Psychol Rev 111:455–485

Pearl J (2000) Causality: models, reasoning & inference. Cambridge University Press, Cambridge, MA

Penn DC, Holyoak KJ, Povinelli DJ (2008) Darwin's mistake: explaining the discontinuity between human and nonhuman minds. Behav Brain Sci 31:109–178

Povinelli DJ (2000) Folk physics for apes. Oxford University Press, Oxford, England

Reips U-D, Waldmann MR (2008) When learning order affects sensitivity to base rates: challenges for theories of causal learning. Exp Psychol 55:9–22

Satpute AJ, Fenker D, Waldmann MR, Tabibnia G, Holyoak KJ, Lieberman M (2005) An fMRI study of causal judgments. Eur J Neurosci 22:1233–1238

Shanks DR, Dickinson A (1987) Associative accounts of causality judgment. In: Bower GH (ed) The psychology of learning and motivation: Advances in research and theory, vol 21. Academic, New York, pp 229–261

Sloman SA (2005) Causal models: how we think about the world and its alternatives. Oxford University Press, Oxford

Sloman SA, Lagnado DA (2005) Do we "do"? Cogn Sci 29(1):5–39

Spirtes P, Glymour C, Scheines P (1993) Causation, prediction & search. Springer, New York

Tenenbaum JB, Griffiths TL (2003) Theory-based causal inference. In: Leen TK, Dietterich TG, Tresp V (eds) Advances in neural information processing systems 15. MIT, Cambridge, MA, pp 35–42

Tomasello M, Call J (1997) Primate cognition. Oxford University Press, Oxford

Waldmann MR (1996) Knowledge-based causal induction. In Shanks D, Holyoak K, Medin D (eds) The psychology of learning and motivation, vol 34. Causal learning. Academic, San Diego, pp 47–88

Waldmann MR (2000) Competition among causes but not effects in predictive and diagnostic learning. J Exp Psychol Learn Mem Cogn 26:53–76

Waldmann MR (2001) Predictive versus diagnostic causal learning: evidence from an overshadowing paradigm. Psychol Bull Rev 8:600–608

Waldmann MR (2007) Combining versus analyzing multiple causes: how domain assumptions and task context affect integration rules. Cogn Sci 31:233–256

Waldmann MR, Hagmayer Y (2001) Estimating causal strength: the role of structural knowledge and processing effort. Cognition 82:27–58

Waldmann MR, Hagmayer Y (2005) Seeing vs. doing: two modes of accessing causal knowledge. J Exp Psychol Learn Mem Cogn 31:216–227

Waldmann MR, Holyoak KJ (1992) Predictive and diagnostic learning within causal models: asymmetries in cue competition. J Exp Psychol Gen 121:222–236

Waldmann MR, Walker JM (2005) Competence and performance in causal learning. Learn Behav 33:211–229

Waldmann MR, Holyoak KJ, Fratianne A (1995) Causal models and the acquisition of category structure. J Exp Psychol Gen 124:181–206

Waldmann MR, Hagmayer Y, Blaisdell AP (2006) Beyond the information given: causal models in learning and reasoning. Curr Dir Psychol Sci 15:307–311

Waldmann MR, Cheng PW, Hagmayer Y, Blaisdell AP (2008) Causal learning in rats and humans: a minimal rational model. In Chater N, Oaksford M (eds) The probabilistic mind. Prospects for Bayesian Cognitive Science. Oxford University Press, Oxford, pp 453–484

White P (2006) The role of activity in visual impressions of causality. Acta Psychol 123:166–185

Woodward J (2003) Making things happen. A theory of causal explanation. Oxford University Press, Oxford

Conditionals: Their Meaning and Their use in Reasoning

Klaus Oberauer

Abstract This chapter gives an overview of the psychology of "if…then", looking at how people understand conditional statements and how they use them in reasoning. There are presently two main theories in the field: The theory of mental models assumes that people interpret conditionals by building mental models of states of affairs meeting their truth conditions, and reason from them by manipulating mental models. The suppositional theory, in contrast, argues that "if…then" expresses a probabilistic relationship between two statements, and assumes that reasoning with conditionals is probabilistic. Experiments on how people interpret conditionals support the suppositional view, whereas experiments on reasoning support the mental-model view, or a dual-process view combining aspects of both theories.

1 Introduction

Conditionals, that is sentences of the form "if … then …", are probably the most important means for expressing our beliefs about how the elements of our world are joined together. We use them to denote causal relations (e.g., "If you take this pill, your headache will go away") and diagnostic ones (e.g., "If the litmus paper turns red, the liquid is acid"), observed regularities (e.g., "If the moon is full, the weather will change"), as well as normative rules (e.g., "If someone helps you out, you should return the favor"), to name just a few. Moreover, conditionals have a prominent role in our reasoning. For instance, they are used to formulate and test scientific hypotheses (e.g., "If my theory is correct, then the treatment should be effective. The treatment was not effective. Therefore, my theory is probably wrong") and decide about actions (e.g., "If I invest in this business, I will make a large profit. Therefore, I will put my money into it"). The meaning of "if" and its use in reasoning seem to be extremely versatile and elusive. It is not surprising,

K. Oberauer
Department of Experimental Psychology, University of Bristol, 12A Priory Road, Bristol BS8 1TU, United Kingdom
e-mail: k.oberauer@bristol.ac.uk

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_9,

therefore, that this little word has attracted much attention from philosophers, linguists, and psychologists who try to pin down what we mean by conditional statements, and to formalize or at least explain its function in sound inductive and deductive reasoning (Bennett J (2003) A philosophical guide to conditionals. Oxford University Press, Oxford; Braine MDS, O'Brien DP (1991) A theory of if: A lexical entry, reasoning program, and pragmatic principles. Psychol Rev 98: 182–203; Evans JSBT, Over DE (2004) If. Oxford University Press, Oxford; Johnson-Laird PN, Byrne RMJ (2002) Conditionals: a theory of meaning, pragmatics, and inference. Psychol Rev 109: 646–678; Oaksford M, Chater N (2001) The probabilistic approach to human reasoning. Trends Cogn Sci 5: 349–357).

The purpose of this chapter is to give an overview of current psychological research on how people untrained in formal logic understand conditionals and how they use conditionals in reasoning. Currently, the field is dominated by two powerful but mutually contradictory theoretical approaches. One is the theory of mental models (Johnson–Laird and Byrne 2002); the other is the probabilistic account, which has been forcefully advanced in the philosophy of logic (e.g., Adams EW (1987) On the meaning of the conditional. Philos Top XV: 5–21; Edgington D (1995) On conditionals. Mind 104: 235–329; Stalnaker R (1991) A theory of conditionals. In: Jackson F (ed) Conditionals. Oxford University Press, Oxford, pp. 28–45), and has engendered several related branches of recent theorizing in psychology (Evans and Over 2004; Liu I (2003) Conditional reasoning and conditionalization. J Exp Psychol Learn, Mem, and Cogn 29: 694–709; Oaksford and Chater 2001). In the remainder of this chapter, I discuss the theory of mental models and one probabilistic theory, with a first pass covering their accounts of the meaning of conditionals and a second pass focusing on their theories of inference.

2 The Meaning of Conditionals

2.1 Mental Models

The theory of mental models (Johnson–Laird and Byrne 1991) has been the dominant theory in the psychology of deductive reasoning for at least a decade. Its basic assumption is that people represent the meaning of descriptive statements, including premises of deductive arguments, as sets of mental models of the possible situations they refer to. The core meaning of a conditional of the form "If p then q" is a set of three mental models, each standing for a possible conjunction of the truth or falsity (denoted by the negation symbol ¬) of the propositions p and q:

p	q
$\neg p$	q
$\neg p$	$\neg q$

For instance, the meaning of "if you take this pill, your headache will go away" is represented by models of the following three situations:

Table 1 Truth tables for different interpretations of conditionals

p	q	$p \supset q$	$p \equiv q$	$p \rightarrow q$
T	T	T	T	T
T	F	F	F	F
F	T	T	F	I
F	F	T	T	I

Note: p and q stand for elementary propositions, T = true, F = false, I = irrelevant, $\supset$ is the material implication, $\equiv$ is the material equivalence or biconditional, and $\rightarrow$ represents the conditional according to the probabilistic view (Adams 1987)

I take the pill, and my headache goes away
I don't take the pill, and my headache goes away
I don't take the pill, and my headache stays

In most circumstances people are assumed to use a sparser representation consisting of a single explicit model of the p & q conjunction (in the example, the conjunction of taking the pill and getting rid of the headache), together with an implicit model. The implicit model (often denoted by "…") stands for further possibilities (e.g., what happens when I don't take the pill) to be elaborated when necessary.

The fully elaborated set of models represents the truth conditions of the *material conditional* (usually denoted as $p \supset q$), which is defined to be true in all cases of the truth table except the case where the antecedent is true and the consequent is false, p & $\neg q$ (see Table 1). The core meaning of *basic* conditionals, according to the theory of mental models, is the material implication. By basic conditionals Johnson-Laird and Byrne (2002) mean conditionals devoid of context and with a content that gives no clues toward a particular interpretation, such as "If there is an A, there is a 2".

The meaning of nonbasic conditionals can be modulated by their content and context. Semantic and pragmatic modulation is assumed to add or subtract models of possibilities (Johnson-Laird and Byrne 2002). For instance, a conditional such as "If you clean my windows, then I will wash your car" pragmatically implies a meaning that allows for only two possibilities, *clean* & *wash* or ¬*clean* & ¬*wash*. The third possibility, that the speaker washes the listener's car even though the listener has not cleaned the speaker's windows, is logically possible but excluded by our knowledge of the pragmatics of promises. A representation of conditionals by these two models corresponds to the *material equivalence* or biconditional in truth-functional semantics (see Table 1).

The material conditional has come under attack as an interpretation of ordinary-language conditionals, first by arguments from philosophy (for a review see Bennett 2003) and later by theoretical and empirical arguments from psychology (Evans et al. 2005). Conceptually, one of the drawbacks of the material implication is that it licenses two paradoxical inferences:

P1: From $\neg p$ we can infer $p \supset q$
P2: From q we can infer $p \supset q$

These inferences are valid because cases with $\neg p$ in the truth table, as well as cases with q, are all cases where $p \supset q$ is true (see Table 1). They are paradoxical because there are examples for which they lead to absurdity if we read $p \supset q$ as "If p then q". Insert, for example, "Egypt is covered in snow throughout the winter" for p and "Palm trees grow on the banks of the Nile" for q.

The inadequacy of the material conditional as a formalization of "If … then" statements motivated the search for alternatives, most of which derive from a footnote of Ramsey (1990), according to which people determine whether to believe "If p then q" by "adding p hypothetically to their stock of knowledge, and arguing on that basis about q; ... they are fixing their degrees of belief in q given p" (p. 155). This note inspired what is sometimes called *the Equation* (Edgington 1995): The rational degree of belief in "If p then q" equals the conditional probability of q, given p. Since "degree of belief" is arguably the same as subjective probability, the Equation states that $P(q|p)$ is the most reasonable value for the subjective probability of the conditional "If p then q".

Evans and Over (2004) endorsed this view and incorporated it into their suppositional theory of conditionals. The theory rests on three assumptions: First, the linguistic function of "if" is to trigger hypothetical thinking, starting from the supposition that the antecedent is true. A conditional "If p then q" makes the claim that q is true under the supposition that p is true. People use the *Ramsey test* for assessing the degree of belief that is warranted in a conditional. The Ramsey test involves estimating the conditional probability of the consequent, given the antecedent, and taking it as the probability of the conditional. For example, if we want to determine how probable we should find the claim "If you take this pill, then your headache will go away", we hypothetically assume that we take the pill, and try to estimate the probability that, in this scenario, our headache goes away.

The second assumption of Evans and Over (2004) is that the meaning of a conditional statement can be enriched by pragmatic implicatures. Given "If p then q", pragmatic and semantic considerations may induce listeners to add the converse "If q then p", or the inverse "If not p then not q", or both. For example, when a speaker says "If you clean my windows, then I will wash your car", the listener is likely to add the inverse, "If you don't clean my windows, then I won't wash your car". She might not add the converse, "If I wash your car, then you will clean my windows". The example illustrates an important difference between the suppositional theory and the model theory. In the suppositional theory, adding the inverse and adding the converse can be independent, whereas in the model theory, they both result from representing the conditional as a biconditional.

Third, Evans and Over (2004) subscribe to a dual-process theory of reasoning, assuming that reasoning is subserved by two systems (cf. Sloman 1996; Stanovich and West 2000). System 1 is characterized as heuristic, fast, automatic, and knowledge based. System 2 is described as analytic, slow, resource demanding, and rule based. Pragmatic implicatures are a result of interpretative processes in System 1; when instructed to reason strictly on the basis of the information given, people can use System 2 to suppress them. I will come back to the dual-process assumption when discussing reasoning from conditionals.

2.2 The Evidence

The probabilistic view clashes with the mental-model theory on a fundamental issue. According to the model theory, the probability of any statement is the probability that one of the possibilities represented by the models for that statement comes true (Johnson-Laird, Legrenzi, Girotto, Legrenzi and Caverni 1999). For any set of mental models this probability, P(MM), is the proportion of all possible cases of the truth table that match one of the models in that set. For instance, the probability of "If you take this pill, your headache will go away" is the probability that one of the three situations represented by the mental models listed in the introduction will come to pass. It can be shown (Lewis 1976) that no probability P(MM) defined in this way is systematically equal to the conditional probability $P(q|p)$. An intuitive way to understand this is by noticing that $P(q|p)$ depends only on the probabilities of the first two cases of the truth table, whereas P(MM) depends on the probability distribution of all four cases. One can, therefore, hold $P(q|p)$ constant – by holding the ratio of p & q to p & $\neg q$ constant – and vary P(MM) by increasing or decreasing the relative probability of the $\neg p$ cases. For instance, P(MM) for the headache conditional can be increased simply by me deciding not to take the pill – this makes it very likely that one of two situations comes true, me not taking the pill and getting rid of the headache, and me not taking the pill and keeping the headache. Both these scenarios render the conditional true, so the conditional ends up with a very high probability. At the same time, my decision not to take the pill has no bearing on the conditional probability that my headache will go away, given that I take the pill, $P(q|p)$. Thus, under the probabilistic view, the probability of the conditional does not change. This means that the probability of a conditional according to the probabilistic account can be dissociated from the probability of the same conditional according to the mental-models account.

This is the logic of a recent series of experiments with the probabilistic truth table task (Evans et al. 2003; Oberauer and Wilhelm 2003). In this task participants are asked to estimate the probability of a given conditional, or to judge whether the conditional is true, in light of information about the frequencies of the four truth-table cases. For instance, they had to evaluate the statement "If the card has a unicorn then it is red", knowing that there are 100 cards with red unicorns, 100 cards with blue unicorns, and 900 cards each of red and blue dragons. This paradigm can be used to dissociate the conditional probability $P(q|p)$ from probabilities of various sets of mental models, including the representation of conditionals by a single p & q model, the biconditional version with two models, and the material-implication version with three models. The critical manipulation is to vary the ratio of p cases to $\neg p$ cases independent of the ratio of p & q to p & $\neg q$. The probabilistic view predicts that people's degree of belief in the conditional depends only on the latter ratio. The mental-model theory predicts that it depends on the ratio of cases matching the mental models to all cases, including the $\neg p$ cases.

Across several experiments with several thousand participants (Oberauer et al. 2005, 2007a, b; Oberauer and Wilhelm 2003) my colleagues and I found that the majority of participants relied exclusively on the ratio of p & q to p & $\neg q$ for their

judgments, providing strong support for the probabilistic view. Judgments of probability of the conditionals were typically very close to the actual value of $P(q|p)$, in line with the Equation. In most of our experiments, however, there was also a minority group who made their judgments depending largely, sometimes exclusively, on the probability of the p & q case (for equivalent findings see Evans et al. 2003). This *conjunctive* response pattern is in line with the assumption of the mental-model theory that people usually represent the conditional by a single explicit model of the p & q conjunction, and that they take only explicit models into account when estimating the probability of a statement (Johnson-Laird et al. 1999).

Details of one of our experiments (Oberauer et al. 2007a, Experiment 3), however, question the mental-model account of the conjunctive response pattern and suggest an alternative explanation in terms of a Ramsey test gone wrong: When conducting the Ramsey test, people initially focus on the p & q case. The correct Ramsey test consists of setting the probability of this case in relation to the probability of all true-antecedent cases, that is, $P(\text{conditional}) = P(q|p) = P(p \,\&\, q)/P(p)$. The incorrect Ramsey test sets $P(p \,\&\, q)$ in relation to all cases, so it computes $P(\text{conditional}) = P(p \,\&\, q)$. For instance, when conducting the Ramsey test on the headache conditional, a person would think "how often in the past have I taken the pill and got rid of my headache?" (p & q), obtaining an estimate of, say, 10 times. The correct Ramsey test would then continue with an estimate of how often in the past the person has taken the pill and the headache has not gone away (p & $\neg q$), obtaining an estimate of, say, 5 times. This yields an estimate of $P(q|p) = 10/15 = 2/3$. A sloppy thinker, however, might replace the second step by an estimate of how often in the past he had a headache (i.e., 100 times), and then compute the probability of the conditional as the probability that the p & q event occurred whenever he had a headache, which is $P(p \,\&\, q) = 10/100 = 1/10$. If that explanation is accepted, the responses of all participants are compatible with the probabilistic view.

3 Inferences from Conditionals

Most psychological research on reasoning from conditionals as premises focuses on four simple inference forms, modus ponens (MP), acceptance of the consequent (AC), denial of the antecedent (DA), and modus tollens (MT), schematically displayed in Table 2. An example for an MP inference is: "If the substance is acid, then the litmus paper will turn red. The substance is acid. Therefore, the litmus paper will turn red". An example for AC starts from the same conditional premise but continues

Table 2 The four basic inference forms

Inference form	Major premise	Minor premise	Conclusion
Modus ponens (MP)	If p then q	p	q
Acceptance of the consequent (AC)	If p then q	q	p
Denial of the antecedent (DA)	If p then q	not p	not q
Modus tollens (MT)	If p then q	not q	not p

Note: The variables p and q stand for elementary propositions; p is called the antecedent and q is called the consequent

with the minor premise "The litmus paper turned red", and concludes "The substance is acid". DA uses the minor premise "The substance is not acid", and the conclusion "The litmus paper will not turn red". Finally, MT has the minor premise "The litmus paper did not turn red", and the conclusion "The substance is not acid".

MP and MT are valid inferences according to formal logic and also according to the suppositional theory. AC and DA are valid only when the conditional premise is interpreted as a biconditional (a plausible interpretation for the acid-litmus example, but not so much for other conditionals). In experiments with basic conditionals, MP is nearly always accepted (or spontaneously generated), whereas about 30–40% of adult participants usually don't accept MT as valid. AC is accepted by about 60% of participants, and DA by slightly fewer (for reviews see Evans 1993; Schroyens et al. 2001).

3.1 Mental Models

The theory of mental models describes deductive reasoning as a three-step procedure. In the first step reasoners construct a set of mental models that meet the truth conditions of all premises. For the argument forms of Table 2 this involves constructing a mental model of the conditional and integrating a model of the minor premise with it. The second step consists of formulating a provisional conclusion that is true in all models of the premises. The third step is a search for counterexamples, that is, possibilities that are compatible with the premises, but not with the conclusion. This means looking for additional mental models which meet the truth conditions of the premises.

The initial representation of a conditional "if p then q" consists of a single explicit mental model of the p & q case, together with an implicit model:

p	...	q

Integrating the positive minor premises of MP and AC into this model is easy – they are already part of the explicit model. The reasoner simply eliminates the implicit model (i.e., the three dots), and thereby gains certainty that the explicit model is the only possible situation. A provisional conclusion can therefore immediately be read off the model. To integrate the negative minor premises of DA and MT with the initial representation of the conditional, reasoners must first flesh out the implicit model, constructing the $\neg p$ & $\neg q$ model, and possibly also the $\neg p$ & q model, thus arriving at two or three explicit models:

p	q
$\neg p$	q
$\neg p$	$\neg q$

Building and holding three models in working memory is assumed to be difficult and error prone. This explains why endorsement rates for MT are lower than for MP, and those for DA are lower than for AC. In the third step of the reasoning process, people should look for counterexamples by fleshing out the implicit model.

This would ideally lead to the construction of an explicit model of $\neg p$ & q. This model is compatible with the premises of AC and DA, but falsifies their conclusions. For example, let's consider an AC inference with the two premises:

1. If my theory is correct, then the treatment should be effective
2. The treatment is effective

Therefore, my theory is correct.

A careful and selfcritical scientist would at this point wonder whether there is an alternative explanation for the outcome of her experiment, and might discover the possibility that her theory is wrong ($\neg p$) and yet the treatment is effective (q) for reasons that have nothing to do with her theory. Considering this possibility comes down to acknowledging that the conclusion of an AC inference, though sometimes highly plausible, is not necessarily true and therefore not deductively valid.

The search for counterexamples explains the difference in acceptance rates between the logically valid inference forms (MP and MT) and the other two forms. Fleshing out the implicit model is difficult because of limits in working memory capacity, and this explains why reasoners often don't succeed in constructing all possible models that are compatible with the premises.

When nonbasic conditionals are used as premises in reasoning, strong content effects can be observed. Research with naturalistic causal conditionals has revealed that acceptance or rejection of an inference is determined mainly by whether people can retrieve a counterexample to the conclusion from their knowledge about the content matter of the argument, and much less by logical form (Cummins 1995; De Neys et al. 2003). For instance, the scientist in our example above might have heard of a colleague's theory that also predicts that the treatment should be effective, and this knowledge makes it more likely for her to consider the $\neg p$ & q possibility. Markovits and Barrouillet (2002) as well as Schroyens and Schaeken (2003) have therefore extended the notion of the search for counterexamples in the mental-model theory. In their versions of the theory, searching for counterexamples means not only trying to construct further mental models that are logically consistent with the premises, but also – and primarily – searching for content-specific counterexamples in memory. Knowledge based counterexamples can also include cases that are incompatible with the conditional premise, that is, cases of p & $\neg q$. For example, given a conditional premise "If a plant is fertilized, then it blooms", and a minor premise, "Peter fertilized his plant", people might retrieve a memory of a well-fertilized plant that died because someone (maybe Peter) neglected to water it. Consequently, they would reject the logically valid MP.

3.2 The Probabilistic View

The probabilistic theories of reasoning with conditionals start from the assumption that our beliefs come in degrees, which can be described as subjective probabilities.

Drawing an inference therefore comes down to deriving a reasonable degree of belief in a conclusion from the subjective probabilities of the premises (Evans and Over 2004; Oaksford et al. 2000).

In the suppositional theory of Evans and Over (2004), reasoning is based on two systems of cognition. The "heuristic" System 1 is assumed to generate MP directly from any given conditional. The conclusion derived by any inference comes with a degree of belief that depends on the degree of belief associated with the premises. This explains why people sometimes reject MP when they have low confidence in the conditional. Experiments have shown that acceptance rates of MP decline when people can retrieve many counterexamples to its conclusion (i.e., cases of $p \,\&\, \neg q$), which also constitute counterexamples to the conditional itself (Cummins 1995; Thompson 1994). The effect of retrieving a counterexample, according to the suppositional theory, is to lower the degree of belief in the conditional premise. For instance, when confronted with the MP inference "If a plant is fertilized, then it blooms. Peter fertilized his plants", System 1 would immediately trigger the conclusion "Peter's plants bloom", unless one or several counterexamples are retrieved from memory of cases in which fertilized plants failed to bloom. These counterexamples would lower the estimate of the conditional probability that plants bloom, given they have been fertilized, and thereby reduce the certainty in the conditional premise, and in the conclusion drawn from it.

The "heuristic" System 1 is assumed to be limited to MP inferences, but this does not limit reasoning as much as it seems, because System 1 is liable to adding the converse and the inverse to a given conditional through pragmatic implicature, and it can even go as far as adding the contrapositive, "If not q then not p". Applying MP to the converse, "If q then p", comes down to endorsing the AC inference for the original conditional. Applying MP to the inverse, "If not p then not q", results in endorsing DA for the original conditional, and applying MP to the contrapositive effectively yields MT. The limitation of System 1 is not so much that it cannot generate inferences but that it cannot discriminate between those that are logically warranted (MP and MT) and those that are not (AC and DA). The "analytical" System 2 contributes to this. On the one hand, it serves to hold pragmatic implicatures at bay, in particular when strictly deductive reasoning is called for. For instance, our scientist can use her System-2 powers to prevent herself from jumping to conclusions, because System 2 would not admit belief in the converse of her initial assumption, "if the treatment is effective, then my theory is true". On the other hand, System 2 provides reasoning strategies, the most important of which is the suppositional strategy to derive MT (originally introduced in the context of rule theories, Braine and O'Brien 1991): Given "If p then q" and "not q", reasoners suppose that p is true, derive from this that q must be true via MP, and notice a contradiction with the minor premise, from which they infer that the supposition must be false and "not p" true. For instance, assume that our scientist has found that her treatment was not effective. She can now reason: "Suppose my theory is true. In that case, the treatment must be effective. In fact, it is not. Thus, the supposition must be false, and hence, my theory must be wrong".

3.3 The Evidence: Reasoning from Conditionals

There is not enough space here to review all relevant evidence on people's reasoning from conditionals, so I'll focus on two domains: Inference acceptance rates for basic conditionals, and content effects.

3.3.1 Patterns of Inference Endorsement

Schroyens and Schaeken (2003) formalized the mental-model theory and fitted it to mean acceptance rates of the four inference forms in Table 2. The model theory provided a good fit, better than the model of Oaksford et al. (2000) that is based on the probabilistic view of conditionals; for a rejoinder see Oaksford and Chater (2003). Unfortunately, the mean acceptance rates of the four inference forms provide only four data points, hardly a strong constraint for any theory. I made an attempt to make use of more information from people's evaluations of the four inference forms by categorizing patterns across the four evaluations (Oberauer 2006). For example, accepting MP, rejecting AC, rejecting DA, and accepting MT is coded as pattern 1001. I fitted multinomial models (Batchelder and Riefer 1999) to the frequencies of the 16 possible patterns obtained from two large data sets, one involving basic conditionals such as "If the square is red, the circle is white", the other involving contextualized causal and noncausal conditionals with fictitious objects, such as "If the Pherotelia blooms, there are blue-point beetles on it". The seven models I tested represented variants of the mental-model theory, of the suppositional theory, and of the theory of Oaksford et al. (2000).

Models incorporating the suppositional reasoning strategy for System 2 (Evans and Over 2004) were not very successful, and models based on the theory of Oaksford et al. (2000) gave a poor account of the data. Of the two models that gave a successful account of the data, one was a version of the mental-model theory that included a parameter for directionality of the models. This parameter captures the assumption that mental models for conditionals have an inherent directionality from the antecedent to the consequent, which makes "forward" inferences (MP and DA) easier than "backward" inferences (AC and MT). This idea was originally proposed by Evans (1993) and empirically supported in later studies (Barrouillet et al. 2000; Oberauer et al. 2005).

The second successful model was a dual-process model which used the assumptions of the suppositional theory to specify the processes of System 1 (i.e., MP inferences and pragmatic implicatures) and the assumptions from the mental-model theory to specify processes of System 2 (i.e., construction of a second model to represent $\neg p$ & $\neg q$, and searching for knowledge-based counterexamples). The blueprint for this model was the dual-process theory by Verschueren et al. (2005), which combines a probabilistic System 1 with a mental-model based System 2.

3.3.2 Content Effects

The model theory and the suppositional theory offer competing explanations of content effects on people's willingness to accept MP and MT. The difference between the two theories is that the effect of counterexamples retrieved from memory on the rejection of conclusions is assumed to be direct in the model theory, whereas according to the suppositional theory it is assumed to be mediated through the subjective $P(q|p)$ and the degree of belief in the conditional premise. My colleagues and I tested these two hypotheses in a path model that allows for a direct as well as a mediated effect (Weidenfeld et al. 2005). The path model was fitted to correlations between the critical variables – availability of counterexamples in memory, ratings of $P(q|p)$, degree of belief in the conditional, and acceptance rates for MP and MT – across 52 contextualized causal and noncausal conditionals. Both the mediated and the direct effects accounted for unique amounts of variance in acceptance rates for MP and MT, but the direct effects were noticeably stronger. This finding is strong evidence that model based reasoning plays an important role in simple inferences from conditionals.

4 Conclusion

This very brief overview of current psychological research on conditionals has shown that two incompatible theoretical approaches, the mental-model theory and the probabilistic view, have complementary strengths and weaknesses. The probabilistic view explains well people's interpretation of conditionals as reflected in truth-table tasks. The mental-model theory has been more successful in explaining people's reasoning with conditionals. Future research will have to find a way to integrate the strengths of both theories.

References

Adams EW (1987) On the meaning of the conditional. Philos Top 15:5–21
Barrouillet P, Grosset N, Lecas JF (2000) Conditional reasoning by mental models: chronometric and developmental evidence. Cognition 75:237–266
Batchelder WH, Riefer DM (1999) Theoretical and empirical review of multinomial process tree modeling. Psychon Bull Rev 6:57–86
Bennett J (2003) A philosophical guide to conditionals. Oxford University Press, Oxford
Braine MDS, O'Brien DP (1991) A theory of if: a lexical entry, reasoning program, and pragmatic principles. Psychol Rev 98:182–203
Cummins DD (1995) Naive theories and causal deduction. Mem Cognit 23:646–658
De Neys W, Schaeken W, d'Ydewalle G (2003) Inference suppression and semantic memory retrieval: every counterexample counts. Mem Cognit 31:581–595
Edgington D (1995) On conditionals. Mind 104:235–329
Evans JSBT (1993) The mental model theory of conditional reasoning: critical appraisal and revision. Cognition 48:1–20

Evans JSBT, Handley SJ, Over DE (2003) Conditionals and conditional probabilities. J Exp Psychol Learn Mem Cogn 29:321–335
Evans JSBT, Over DE (2004) If. Oxford University Press, Oxford
Evans JSBT, Over DE, Handley S (2005) Suppositions, extensionality, and conditionals: A critique of the mental-models theory of Johnson-Laird and Byrne (2002). Psychol Rev 112:1040–1052
Johnson-Laird PN, Byrne RMJ (1991) Deduction. Erlbaum, Hillsdale
Johnson-Laird PN, Byrne RMJ (2002) Conditionals: a theory of meaning, pragmatics, and inference. Psychol Rev 109:646–678
Johnson-Laird PN, Legrenzi P, Girotto V, Legrenzi MS, Caverni JP (1999) Naive probability: a mental model theory of extensional reasoning. Psychol Rev 106:62–88
Lewis D (1976) Probabilities of conditionals and conditional probabilities. Philos Rev 85:297–315
Liu I (2003) Conditional reasoning and conditionalization. J Exp Psychol Learn Mem Cogn 29:694–709
Markovits H, Barrouillet P (2002) The development of conditional reasoning. Dev Rev 22:5–36
Oaksford M, Chater N (2001) The probabilistic approach to human reasoning. Trends Cogn Sci 5:349–357
Oaksford M, Chater N (2003) Computational levels and conditional inference: reply to schroyens and schaeken (2003). J Exp Psychol Learn Mem Cogn 29:150–156
Oaksford M, Chater N, Larkin J (2000) Probabilities and polarity biases in conditional inference. J Exp Psychol Learn Mem Cogn 26:883–899
Oberauer K (2006) Reasoning with conditionals: a test of formal models of four theories. Cogn Psychol 53:238–283
Oberauer K, Geiger SM, Fischer K, Weidenfeld A (2007a) Two meanings of "if"? Individual differences in the interpretation of conditionals. Q J Exp Psychol 60:790–819
Oberauer K, Weidenfeld A, Fischer K (2007b) What makes us believe a conditional? The roles of covariation and causality. Thinking & Reasoning, 13:340–369
Oberauer K, Hörnig R, Weidenfeld A, Wilhelm O (2005) Effects of directionality in deductive reasoning, II: Premise integration and conclusion evaluation. Q J Exp Psychol 58A:1225–1247
Oberauer K, Wilhelm O (2003) The meaning(s) of conditionals – Conditional probabilities, mental models, and personal utilities. J Exp Psychol Learn Mem Cognit 29:680–693
Ramsey FP (1990) General propositions and causality (originally published 1929). In: Mellor DH (ed) Philosophical papers F. P. Ramsey. Cambridge University Press, Cambridge, pp 145–163
Schroyens W, Schaeken W (2003) A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning. J Exp Psychol Learn Mem Cognit 29:140–149
Schroyens W, Schaeken W, d'Ydewalle G (2001) The processing of negations in conditional reasoning: a meta-analytic case study in mental model and/or mental logic theory. Think Reas 7:121–172
Sloman SA (1996) The empirical case for two systems of reasoning. Psychol Bull 119:3–22
Stalnaker R (1991) A theory of conditionals. In: Jackson F (ed) Conditionals. Oxford University Press, Oxford, pp 28–45
Stanovich KE, West RF (2000) Individual differences in reasoning: implications for the rationality debate? Behav Brain Sci 23:645–726
Thompson VA (1994) Interpretational factors in conditional reasoning. Mem Cognit 22:742–758
Verschueren N, Schaeken W, d'Ydewalle G (2005) A dual-process specification of causal conditional reasoning. Think Reas 11:293–278
Weidenfeld A, Oberauer K, Hörnig R (2005) Causal and noncausal conditionals – an integrated model of interpretation and reasoning. Q J Exp Psychol 58A:1479–1513

Thinking and Memory

Matthias Brand and Hans J. Markowitsch

Abstract Remembering the past is crucially important for cognitive functions, such as anticipating and planning future activities or thinking about one's own self. In Tulving's hierarchy of long-term memory systems, episodic memory is the highest one that is most likely uniquely human. One of the characteristics of episodic memory is the ability to mentally travel into the past and the future. Several brain structures are fundamentally involved in successful acquisition and retrieval of episodic memories. In particular, limbic regions and parts of the prefrontal cortex are associated with specific facets of episodic memory, i.e. processing the emotional connotation of personal experiences and self-relevant information. Additionally, other brain regions have important supportive functions in memory encoding and retrieval. Specifically, the dorsolateral prefrontal cortex is engaged in learning strategies and in metacognitive processes necessary for successfully remembering information stored in long-term memory. Both studies with brain damaged patients and investigations employing functional neuroimaging techniques have provided insights into the neural correlates that associate thinking and memory.

1 Introduction

Thinking and memory are strongly interrelated. Given that thinking comprises several facets of cognition including problem solving, judgement and decision making, self-reflection and anticipating future events on the basis of past experiences, it is obvious to everyone that thinking is mainly influenced by the ability to remember one's own past, the representation of knowledge, and the way we perceive the world based upon perceptual learning processes. On the other side,

M. Brand (✉) and H.J. Markowitsch
Department of General Psychology: Cognition - Unit Applied Cognitive Science - University of Duisburg-Essen - Campus Duisburg - Forsthausweg 2 - 47048 Duisburg, Germany
e-mail: matthias.brand@uni-due.de; hjmarkowitsch@uni-bielefeld.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_10, © Springer-Verlag Berlin Heidelberg 2010

travelling mentally into one's own past, which is one main component of memory, also requires subcomponents of thinking, e.g. reflective processes, associative thinking, executive control, inhibitory processes, and other cognitive functions like selective attention.

Thinking allows an individual to deal with a complex world in accordance with his or her personal desires, plans, goals and beliefs. Memory allows an individual to act in a world that changes from moment to moment and to anticipate future experiences. In other words, thinking and memory normally act in concert. The relationship between thinking and memory (or thoughts and memories) can be found on both behavioural and brain levels. In this contribution we will concentrate on the link between thinking and the ability to remember biographical episodes. In particular, we will focus on mental time travelling and self-reflection as major components of autobiographical-episodic memory. These components include several aspects of thinking: thinking about the future, thinking about the own self, thinking about personal relevance of various past, current, or future circumstances and so on. We will demonstrate how these components interact with episodic memory and which brain regions are neural bases for this relationship. We will argue that mental time travelling, in particular, may be understood as a bridge between memory and thinking.

We first introduce some definitions and classifications of memory, and we emphasise the specific characteristics of the highest memory system in Tulving's hierarchy of long-term memory (e.g. Tulving 1995, 2002) which is episodic memory. Thereafter, we summarise brain correlates of episodic memory processes. In the following two sections, we give evidence for the relationship between thinking and memory in patients with brain damage or dysfunctions and common neural correlates of thinking and memory as revealed by functional brain imaging studies. A general conclusion will close the chapter.

2 Definitions and Classifications of Memory

The ability to remember the past is one of the most fascinating phenomena in human beings. Memory research has provided good insights into the cognitive and brain correlates of different memory processes. For research purposes, the distinction between specific types of memory is necessary. For instance, recalling a phone number 20 s after it has been heard is different from remembering a personal event that has happened many years ago. Likewise, the ability to consciously encode ten items that have to be bought 30 min later in a supermarket is different from recognising a familiar melody. Given these few examples of the different features of memory, it is obvious that we need definitions and classifications when describing different functions of memory. Additionally, potential behavioural and brain correlates involved in memory as well as associations between thinking and memory need to be defined.

Memory can be distinguished with respect to time, process, and content. As Atkinson and Shiffrin (1968) already proposed more than 30 years ago, memory can be differentiated on a time axis into short-term and long-term memory. Additionally, working memory, a system that interacts with both short-term and long-term memory, actively binds new information with that stored in the long-term memory system (Baddeley 1986, 2000). Beyond these time-oriented distinctions of memory, there is a further specification of memory functioning along the time axis that involves the ability to mentally travel to the past and future. Mental time travelling is one of the most important features of what is called "episodic memory" (see below).

On a content-based level, memory has been defined differently in the past. Endel Tulving (1972, 1995, 2002) proposed one of the most influencing theories of memory. He initially hypothesised four long-term memory systems which are assumed to be hierarchically organised. The lowest system, the procedural memory system, comprises procedures such as riding a bike or playing cards. Both acquisition and retrieval of procedural memories are assumed to be unconscious and non-verbal. The next system is the priming system in which the unconscious encoding of information results in a higher and successful recognition rate, even when only some details are presented (cf. word-stem completion tasks). Semantic and episodic memory are the two remaining memory systems with both encompassing the conscious acquisition and retrieval of information. While the semantic memory system consists of facts (e.g. world knowledge) and is noëtic, the episodic memory is comprised of biographical events and is therefore autonoëtic. In recent years, a further memory system known as perceptual memory was introduced by Tulving and Markowitsch (cf. Markowitsch 2003b). This system, placed between the priming system and semantic memory, allows individuals to have a feeling of familiarity with an object without explicitly knowing the meaning or the name of that object. The five memory systems are illustrated in Fig. 1.

The definition of "episodic memory" has been changed and shaped since its introduction more than 30 years ago. In his recent definition, Tulving (2002, 2005) accentuates that episodic memory is not restricted to time and space, but is the conjunction of subjective time, autonoëtic consciousness and the experience of self. Moreover, episodic memories usually have an emotional connotation. Given this definition, it is obvious that Tulving stresses episodic memory as a system that is uniquely human. Nonetheless, some animal researchers believe that episodic memory is also present in non-human species – an argumentation that is frequently based on behavioural observations of animals in specific test conditions that presumably measure some kinds of episodic memory (e.g. Babb and Crystal 2006; Clayton et al. 2003; Skov-Rackette et al. 2006) (see also the recent review by Dere et al. 2006). Based on Tulving's definition, animals would have to possess all the specific components that constitute episodic memory (see above); so far, however, there is no convincing evidence suggesting that animals have these abilities (Premack 2007).

In recent years, a debate has arisen about the impact of mental time travelling as a crucial component of episodic memory (Schacter et al. 2007). Mental time travelling

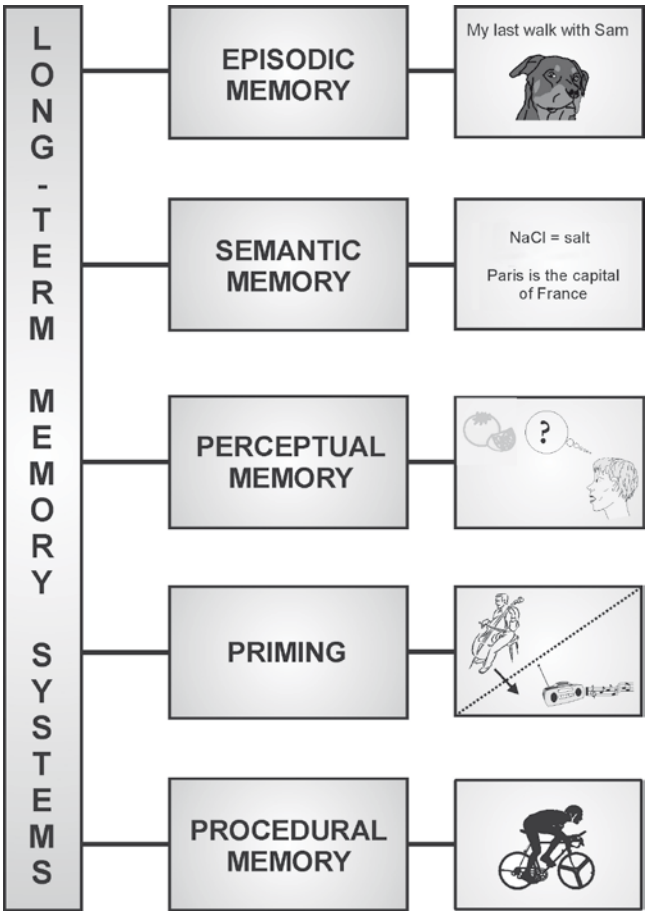

Fig. 1 The five long-term memory systems, as suggested by Tulving and Markowitsch (cf. Markowitsch 2003b)

refers to the "mental activity in which people engage when they remember particular past events, or think about possible personal future happenings" (Tulving and Kim 2007, p. 335). However, other terminologies such as focusing on abilities that allow an individual to anticipate the future are also used (cf. Suddendorf and Corballis 2007). Considering the context of the discussion on mental time travelling, it has also been controversially debated whether the process of mentally travelling to the past or to the future is uniquely human (Tulving 2005; Tulving and Kim 2007), or whether other animals also are capable of performing this mental activity (see also the discussion on episodic-like memory above) (cf. Suddendorf and Corballis 2007). For an intriguing review on human and animal cognition, in which the capacity to plan future actions is also an issue, see the article by Premack (2007).

3 Brain Structures Involved in Episodic Memory: A Brief Summary

Given the distinction between the memory systems described above, it is obvious that several brain systems are engaged in memory processes. Beyond some basal networks necessary for all kinds of memory processes (i.e. those involved in general attention and information processing), several brain regions act as so-called bottleneck-structures with respect to the memory system that is responsible for information that needs to be encoded or retrieved (Brand and Markowitsch 2003). In other words, for the processes of encoding, storing and retrieval of episodic and semantic as well as of perceptual or procedural information, different brain structures play crucial roles (and for the phenomenon of priming as well). In the following sections, we focus on human brain circuits necessary for successful encoding and retrieval of episodic memories, because we assume that the episodic memory system is most likely uniquely human (Tulving 2005), and is mediated by structures which are in part only existent – or particularly developed – in the human species (Markowitsch 1988, 1994, 2000; Markowitsch et al. 1985; Markowitsch and Tulving 1994).

For encoding and consolidation of episodic memories (and at least also partially for semantic memory), two limbic circuits are considered as primary neural correlates. The first circuit, frequently referred to as the Papez circuit (Papez 1937), consists of the hippocampal formation, the mammillary bodies, the anterior thalamic nuclei, and the cingulate gyrus. These structures are interconnected through several fibre tracts, such as the fornix, the mammillothalamic tract, the thalamic pedunculi, and the cingulum (see Figs. 2 and 3). The recent view of these structures and their role in episodic memory is that they are principally engaged in acquiring information, through binding new information to that which is already

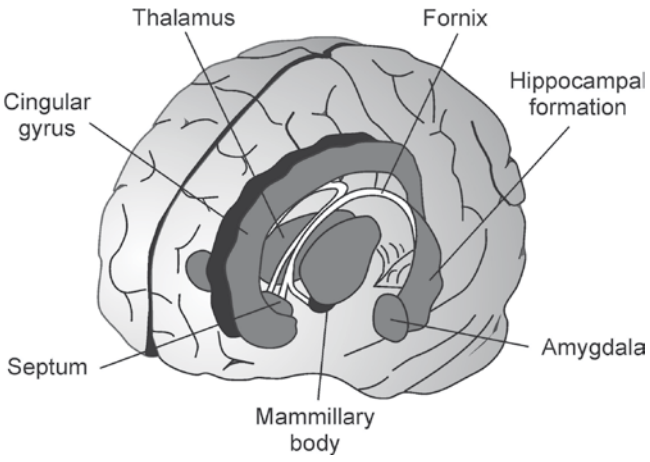

Fig. 2 Structures and fibre tracts of the limbic system

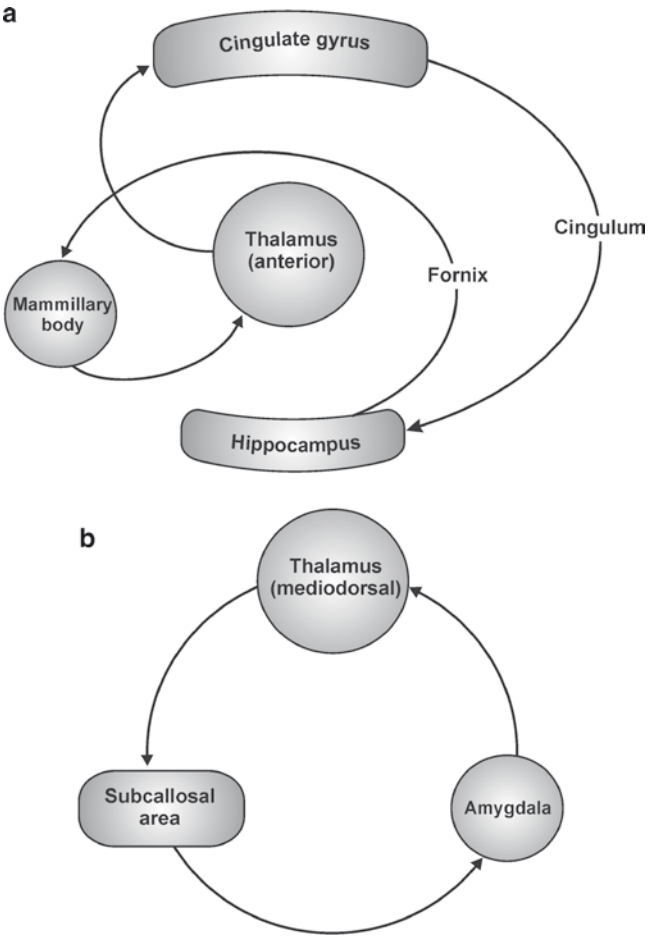

Fig. 3 A schematic illustration of the Papez circuit (**a**) and the basolateral-limbic circuit (**b**)

stored in long-term memory. When information to be encoded is emotionally coloured – which is by definition the case in episodic memories – a second circuit becomes additionally important. This second circuit which is frequently referred to as the basolateral-limbic or amygdaloid circuit consists of the amygdala and surrounding limbic structures as well parts of the orbitofrontal cortex (Fig. 3). Its function is to bind emotionally relevant information during memory building. This is a function that is exceptionally linked to the amygdala (Cahill 2000; Cahill et al. 2001; Fujiwara and Markowitsch 2006; Markowitsch 2000), because the amygdala is the "par excellence" structure for evaluating emotional sensory stimuli (e.g. Phelps 2006; Phelps and LeDoux 2005).

Evidence for the involvement of these circuits in encoding of episodic memories comes from recent neuroimaging studies with healthy volunteers and patient

populations (e.g. Binder et al. 2005; Cabeza et al. 1997, 2002; Cabeza and Nyberg 2000; Cabeza and St Jacques 2007; Greicius et al. 2003; Kircher et al. 2007; Kumaran and Maguire 2006; Nyberg et al. 1996; Rand-Giovannetti et al. 2006; Uncapher and Rugg 2005a,b). Likewise, several studies have also found these structures to be activated during retrieval of episodic memories (Fink et al. 1996; Haist et al. 2001; Levine et al. 2004; Moscovitch et al. 2005; Piefke et al. 2003, 2005; Steinvorth et al. 2005; Svoboda et al. 2006; Vandekerckhove et al. 2005), although their specific role in retrieval processes is still a topic of debate. For instance, it has been reported that the hippocampal formation's contribution to retrieval is moderated by a subject's age or gender, or by the memories' age (Piefke and Fink 2005; Piefke et al. 2005; Viard et al. 2007).

In addition to the limbic circuits, the prefrontal cortex (Fig. 4) is also fundamentally involved in the encoding and retrieval of episodic memories. In particular, the dorsolateral prefrontal cortex is engaged in encoding new information, though its specific role is still a topic of debate (see comments below). The involvement of the dorsolateral section of the frontal lobe in retrieving information stored in long-term memory applies to both episodic and semantic information (Brand and Markowitsch 2008; Vandekerckhove et al. 2005). Its contribution to successful retrieval becomes more crucial as retrieval conditions become difficult and require effort (Buckner 2003; Lepage et al. 2000; Rugg et al. 2002; Velanova et al. 2003) (see comments below). There are also some reports on gender or age effects on the engagement of the dorsolateral prefrontal cortex in remembering the past (Piefke and Fink 2005; Piefke et al. 2005).

Beyond the involvement of the dorsolateral prefrontal region, the orbitofrontal cortex is also engaged in encoding and retrieving episodic memories, primarily

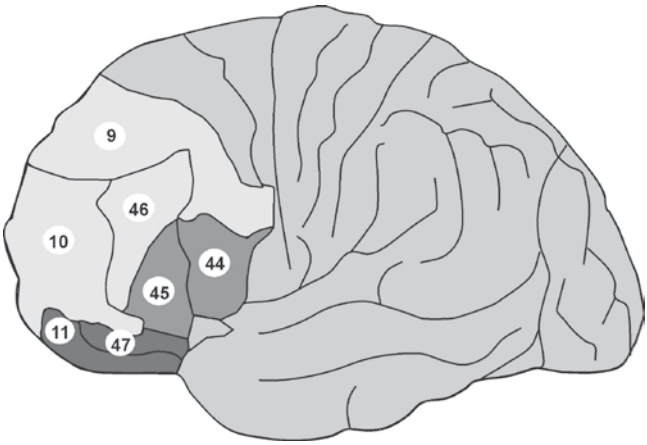

Fig. 4 The different sections of the lateral prefrontal cortex. Numbers indicate Brodmann areas (BA). The dorsolateral prefrontal cortex consists of BA 9, 10, and 46. The orbitofrontal region (also frequently referred to as the inferolateral prefrontal cortex) comprises BA 11 and 47. In addition, for a better orientation the location of Broca's area (parts of BA 44 and 45) is also included in the figure

when information to be learned or to be remembered is highly emotionally coloured (Brand and Markowitsch 2006; Cabeza et al. 2004; Herholz et al. 2001; Markowitsch et al. 2003; Piefke et al. 2003; Svoboda et al. 2006).

4 The Association Between Thinking and Memory in Brain Damaged Patients

In patients with damage to specific brain structures, a dissociation of impaired memory functioning and intact general intellectual abilities is observable. For instance, in patients with medial temporal lobe pathology anterograde amnesia is the most prominent symptom, while intelligence and other intellectual abilities are unaffected (Bird et al. 2007; Corkin 2002; Markowitsch 1992; Scoville and Milner 1957). Other examples representing dissociation of impaired memory, but spared intelligence are patients with alcoholic Korsakoff's syndrome. Such patients have severe anterograde memory deteriorations while maintaining almost normal intelligence (Brand 2007; Kopelman 1995). Nevertheless, it has also been demonstrated that in these patients other functions (e.g. problem solving, reasoning and decision-making) can be affected (e.g. Brand et al. 2005; Brokate et al. 2003). These findings emphasise the view that neuropsychological functions are – at least to a moderate degree – correlated with each other (Markowitsch 1992, 2003a). A high correlation between memory disorders and additional cognitive symptoms can be seen in patients with more general brain dysfunctions, such as in patients with dementia (e.g. Metzler-Baddeley 2007). Figure 5 provides an example of a partial dissociation of memory impairments and other cognitive functions, as well as an example of a neuropsychological profile that represents a more general cognitive decline.

As pointed out earlier in this chapter, a specific component of episodic memory is mental time travelling. In accordance with this position, it has been

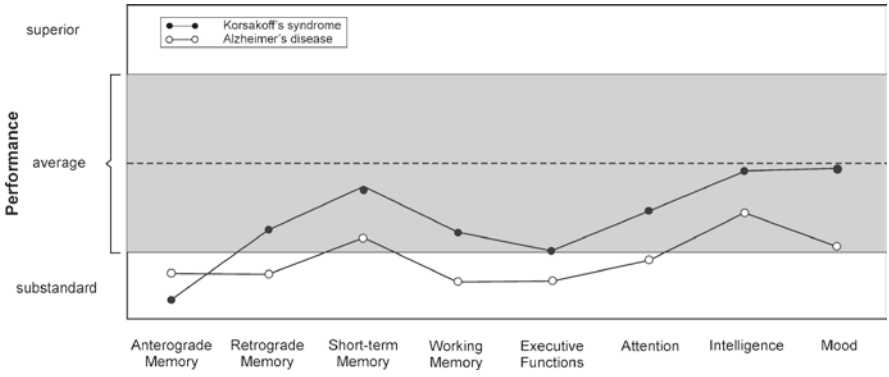

Fig. 5 A neuropsychological profile of a patient with alcoholic Korsakoff's syndrome and that of a patient with Alzheimer's disease in a moderate state of the syndrome

recently demonstrated that patients who are impaired in episodic memory also have difficulties in anticipating future events. In particular, the investigation by Hassabis et al. (2007b) has demonstrated that patients with hippocampal lesions are impaired in imagining future experiences (see also Atance and O'Neill 2001). In the study by Hassabis et al., patients' descriptions of imagined events were less detailed regarding the environmental setting. Specifically, the images of experiences were reduced with respect to spatial coherence. This finding supports the assumption of a strong correlation between the ability to vividly re-experience past events and to actively imagine the future (see also Schacter et al. 2007 for a comprehensive review of this issue). An association between anterograde episodic memory impairments, difficulties in processing time information, and in estimating time intervals has also been observed in patients with Korsakoff's syndrome (Brand et al. 2003; Mimura et al. 2000).

Other relationships between memory and other cognitive functions can be found in patients with schizophrenia. In particular, Corcoran and Frith (2003) have revealed that patients with schizophrenia have difficulties in autobiographical retrieval (their memories lack specificity). The memory reductions were related to impairments in Theory-of-Mind abilities (i.e. understanding thoughts and feelings of others).

5 The Association Between Thinking and Memory in Neuroimaging Investigations

As mentioned above, several brain regions are involved in the encoding and retrieval of episodic memories, although their specific contribution to the encoding or the retrieval process is still unclear. This is particularly the case for the dorsolateral prefrontal cortex. This structure is frequently activated in functional imaging studies that have investigated the neural correlates of encoding and retrieval; however, it remains unclear whether the dorsolateral section is directly linked to memory processes, or whether it supports memory acquisition or remembering through its role in executive functioning. The dorsolateral section of the frontal lobe is strongly associated with working memory, higher order executive functioning, and metacognitive processes (Elliott 2003; Fuster 2006; Kane and Engle 2002; Lie et al. 2006) which are fundamentally important for building and applying encoding strategies. More specifically, executive functions are necessary for organising and categorising new material and for reducing the complexity of information to be learned. In these memory associated functions the dorsolateral prefrontal region plays a key role, at least when the information to be encoded is complex, and the situation requires learning strategies and reflective processes (Buckner 2003; Jansma et al. 2007; Miotto et al. 2006; Ranganath et al. 2003). An example of a learning strategy is the forming of associations between items; it has been shown that the dorsolateral prefrontal cortex supports this process (Addis and McAndrews 2006; Blumenfeld and Ranganath 2006; Staresina and Davachi 2006;

Summerfield et al. 2006). In addition, this prefrontal region seems to be a central neural correlate of processing during general retrieval mode and retrieval effort (Lepage et al. 2000; Velanova et al. 2003). In summary, memory processes and executive functions – or more generally, memory and thinking – are substantially interrelated on a functional level. Most likely, the dorsolateral prefrontal cortex represents an important neural correlate of the association between thinking and memory (cf. Brand and Markowitsch 2008).

Another process fundamentally involved in episodic memory is processing self relevant information. It has been consistently found that the dorsomedial prefrontal cortex, the anterior cingulate gyrus and the retrosplenial cortex are engaged in retrieving autobiographical memories or self-referentially encoded material (e.g. Fossati et al. 2004; Piefke et al. 2003; Svoboda et al. 2006). These findings are consistent with reports about the fundamental contribution of these regions to thinking about one's own self (Johnson et al. 2002; Northoff et al. 2006; Schmitz and Johnson 2006; Schmitz et al. 2004).

We have already introduced the association between mental time travelling and episodic memory and have summarised some neuropsychological evidence for this association. In addition to these findings, recent functional imaging studies with healthy subjects also point to a clear relationship between remembering the past and thinking about the future by revealing widely overlapping neural correlates for these two important functions (see the excellent review by Hassabis and Maguire 2007). In particular, it has been demonstrated by Hassabis et al. (2007a) that the hippocampus, the parahippocampal gyrus and the retrosplenial cortex were activated during both re-experiencing past events and constructing new experiences. Anterior medial prefrontal regions, as well as parts of the parietal cortex and the precuneus were associated with self-schema activation and may support differentiating between true and fictitious events. Hassabis et al. conclude that episodic memory and episodic future thinking are crucially associated and that these functions share some similar brain networks.

6 General Considerations

In summary, memory and thinking are strongly correlated. Although there are naturally some very specific components involved in both memory and thinking, the two functions share several fundamental processes and underlying neural networks (in particular prefrontal and limbic regions). Accordingly, patients with brain pathology who have episodic memory reductions also frequently have difficulties in other cognitive domains. One of the most important functions that combine memory and thinking is mental time travelling or, in other words, thinking about the future. The importance of the ability to anticipate the future on the basis of past experiences has been intriguingly expressed by Schacter et al. (2007, p. 660) in the paragraph entitled "The prospective brain," in which he writes "preparing for the future is a vital task in any domain of cognition or behaviour that is important for

survival." We would like to expand this conclusion by referring to Ewald Hering (1870) who – during his famous talk at the University of Vienna – pointed out the role of memory in cognition as a unifying force that holds the self together.

References

Addis DR, McAndrews MP (2006) Prefrontal and hippocampal contributions to the generation and binding of semantic associations during successful encoding. Neuroimage 33:1194–1206

Atance CM, O'Neill DK (2001) Episodic future thinking. Trends Cogn Sci 5:533–539

Atkinson RC, Shiffrin RM (1968) Human memory: a proposed system and its control processes. In: Spence KW, Spence JT (eds) The psychology of learning and motivation: advances in research and theory, vol 2. Academic Press, New York, pp 89–195

Babb SJ, Crystal JD (2006) Episodic-like memory in the rat. Curr Biol 16:1317–1321

Baddeley AD (1986) Working memory. University Press, Oxford

Baddeley AD (2000) The episodic buffer: A new component of working memory? Trends Cogn Sci 4:417–423

Binder JR, Bellgowan PS, Hammeke TA, Possing ET, Frost JA (2005) A comparison of two FMRI protocols for eliciting hippocampal activation. Epilepsia 46:1061–1070

Bird CM, Shallice T, Cipolotti L (2007) Fractionation of memory in medial temporal lobe amnesia. Neuropsychologia 45:1160–1171

Blumenfeld RS, Ranganath C (2006) Dorsolateral prefrontal cortex promotes long-term memory formation through its role in working memory organization. J Neurosci 26:916–925

Brand M (2007) Cognitive profile of patients with alcoholic Korsakoff's syndrome. Int J Disabil Hum Dev 6:161–170

Brand M, Fujiwara E, Borsutzky S, Kalbe E, Kessler J, Markowitsch HJ (2005) Decision-making deficits of Korsakoff patients in a new gambling task with explicit rules: associations with executive functions. Neuropsychology 19:267–277

Brand M, Kalbe E, Fujiwara E, Huber M, Markowitsch HJ (2003) Cognitive estimation in patients with probable Alzheimer's disease and alcoholic Korsakoff patients. Neuropsychologia 41:575–584

Brand M, Markowitsch HJ (2003) The principle of bottleneck structures. In: Kluwe RH, Lüer G, Rösler F (eds) Principles of learning and memory. Birkhäuser, Basel, pp 171–184

Brand M, Markowitsch HJ (2006) Memory processes and the orbitofrontal cortex. In: Zald D, Rauch S (eds) The orbitofrontal cortex. Oxford University Press, Oxford, pp 285–306

Brand M, Markowitsch HJ (2008) The role of the prefrontal cortex in episodic memory. In: Dere E, Huston JP, Easton A (eds) Handbook of episodic memory. Elsevier, Amsterdam, pp 317–342

Brokate B, Hildebrandt H, Eling P, Fichtner H, Runge K, Timm C (2003) Frontal lobe dysfunctions in Korsakoff's syndrome and chronic alcoholism: continuity or discontinuity? Neuropsychology 17:420–428

Buckner RL (2003) Functional-anatomic correlates of control processes in memory. J Neurosci 23:3999–4004

Cabeza R, Dolcos F, Graham R, Nyberg L (2002) Similarities and differences in the neural correlates of episodic memory retrieval and working memory. Neuroimage 16:317–330

Cabeza R, Grady LC, Nyberg L, McIntosh RA, Tulving E, Kapur S, Jennings MJ, Houle S, Craik MIF (1997) Age-related differences in neural activity during memory encoding and retrieval: A positron emission tomography study. J Neurosci 17:391–400

Cabeza R, Nyberg L (2000) Imaging cognition II: An empirical review of 275 PET and fMRI studies. J Cogn Neurosci 12:1–47

Cabeza R, Prince SE, Daselaar SM, Greenberg DL, Budde M, Dolcos F, LaBar DL, Rubin DC (2004) Brain activity during episodic retrieval of autobiographical and laboratory events: an fMRI study using a novel photo paradigm. J Cogn Neurosci 16:1583–1594

Cabeza R, St Jacques P (2007) Functional neuroimaging of autobiographical memory. Trends Cogn Sci 11:219–227

Cahill L (2000) Modulation of long-term memory in humans by emotional arousal: adrenergic activation and the amygdala. In: Aggleton JP (ed) The amygdala: a functional analysis. Oxford University Press, Oxford, pp 425–446

Cahill L, Haier RJ, White NS, Fallon J, Kilpatrick L, Lawrence C, Potkin SG, Alkire MT (2001) Sex-related differences in amygdala activity during emotionally influenced memory storage. Neurobiol Learn Mem 75:1–9

Clayton NS, Bussey TJ, Dickinson A (2003) Can animals recall the past and plan for the future? Nature Rev Neurosci 4:685–691

Corcoran R, Frith CD (2003) Autobiographical memory and theory of mind: evidence of a relationship in schizophrenia. Psychol Med 33:897–905

Corkin S (2002) What's new with the amnesic patient H.M.? Nature Rev Neurosci 3:153–160

Dere E, Kart-Teke E, Huston JP, De Souza Silva MA (2006) The case for episodic memory in animals. Neurosci Biobehav Rev 30:1206–1224

Elliott R (2003) Executive functions and their disorders. Br Med Bull 65:49–59

Fink GR, Markowitsch HJ, Reinkemeier M, Bruckbauer T, Kessler J, Heiss W-D (1996) Cerebral representation of one's own past: neural networks involved in autobiographical memory. J Neurosci 16:4275–4282

Fossati P, Hevenor SJ, Lepage M, Graham SJ, Grady C, Keightley ML, Craik F, Mayberg H (2004) Distributed self in episodic memory: neural correlates of successful retrieval of self-encoded positive and negative personality traits. Neuroimage 22:1596–1604

Fujiwara E, Markowitsch HJ (2006) Brain correlates of binding processes of emotion and memory. In: Zimmer H, Mecklinger AM, Lindenberger U (eds) Binding in human memory - A neurocognitive perspective. Oxford University Press, Oxford, pp 379–410

Fuster JM (2006) The cognit: a network model of cortical representation. Int J Psychophysiol 60:125–132

Greicius MD, Krasnow B, Boyett-Anderson JM, Eliez S, Schatzberg AF, Reiss AL, Menon V (2003) Regional analysis of hippocampal activation during memory encoding and retrieval: fMRI study. Hippocampus 13:164–164

Haist F, Bowden Gore J, Mao H (2001) Consolidation of human memory over decades revealed by functional magnetic resonance imaging. Nature 4:1139–1145

Hassabis D, Kumaran D, Maguire EA (2007a) Using imagination to understand the neural basis of episodic memory. J Neurosci 27:14365–14374

Hassabis D, Kumaran D, Vann SD, Maguire EA (2007b) Patients with hippocampal amnesia cannot imagine new experiences. Proc Natl Acad Sci USA 104:1726–1731

Hassabis D, Maguire EA (2007) Deconstructing episodic memory with construction. Trends Cogn Sci 11:299–306

Herholz K, Ehlen P, Kessler J, Strotmann T, Kalbe E, Markowitsch HJ (2001) Learning face-name associations and the effect of age and performance: a PET activation study. Neuropsychologia 39:643–650

Hering KEK (1870, 30.05.1870). Über das Gedächtnis als eine allgemeine Function der organisirten Materie. Paper presented at the Kaiserliche Akademie der Wissenschaften, Wien

Jansma JM, Ramsey NF, de Zwart JA, van Gelderen P, Duyn JH (2007) fMRI study of effort and information processing in a working memory task. Hum Brain Mapp 28:431–440

Johnson SC, Baxter LC, Wilder LS, Pipe JG, Heiserman JE, Prigatano GP (2002) Neural correlates of self-reflection. Brain 125:1808–1814

Kane MJ, Engle RW (2002) The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective. Psychon Bull Rev 9:637–671

Kircher TT, Weis S, Freymann K, Erb M, Jessen F, Grodd W, Heun R, Leube DT (2007) Hippocampal activation in patients with mild cognitive impairment is necessary for successful memory encoding. J Neurol Neurosurg Psychiatry 78:812–818

Kopelman MD (1995) The Korsakoff syndrome. Br J Psychiatry 166:154–173

Kumaran D, Maguire EA (2006) The dynamics of hippocampal activation during encoding of overlapping sequences. Neuron 49:617–629
Lepage M, Ghaffar O, Nyberg L, Tulving E (2000) Prefrontal cortex and episodic memory retrieval mode. Proc Natl Acad Sci USA 97:506–511
Levine B, Turner GR, Tisserand D, Hevenor SJ, Graham SJ, McIntosh AR (2004) The functional neuroanatomy of episodic and semantic autobiographical remembering: a prospective functional MRI study. J Cogn Neurosci 16:1633–1646
Lie CH, Specht K, Marshall JC, Fink GR (2006) Using fMRI to decompose the neural processes underlying the Wisconsin Card Sorting Test. Neuroimage 30:1038–1049
Markowitsch HJ (1988) Anatomical and functional organization of the primate prefrontal cortical system. In: Steklis HD, Erwin J (eds) Comparative primate biology, Vol. IV: Neurosciences. Alan R. Liss, New York, pp 99–153
Markowitsch HJ (1992) Intellectual functions and the brain. An historical perspective. Hogrefe and Huber Publs, Toronto
Markowitsch HJ (1994) Brain development and behavior. In: Husen T, Postlethwaite TN (eds) The international encyclopedia of education (Vol 2nd). Pergamon Press, Oxford, pp 552–554
Markowitsch HJ (2000) The anatomical bases of memory. In: Gazzaniga MS (ed) The new cognitive neurosciences, 2nd edn. The MIT Press, Cambridge, pp 781–795
Markowitsch HJ (2003a) Memory: disturbances and therapy. In: Brandt T, Caplan L, Dichgans J, Diener HC, Kennard C (eds) Neurological disorders; course and treatment, 2nd edn. Academic Press, San Diego, pp 287–302
Markowitsch HJ (2003b) Psychogenic amnesia. Neuroimage 20:S132–S138
Markowitsch HJ, Emmans D, Irle E, Streicher M, Preilowski B (1985) Cortical and subcortical afferent connections of the primate's temporal pole: a study of rhesus monkeys, squirrel monkeys, and marmosets. J Comp Neurol 242:425–458
Markowitsch HJ, Tulving E (1994) Cognitive processes and cerebral cortical fundi: Findings from positron-emission tomography studies. Proc Natl Acad Sci USA 91:10507–10511
Markowitsch HJ, Vandekerckhove MM, Lanfermann H, Russ MO (2003) Engagement of lateral and medial prefrontal areas in the ecphory of sad and happy autobiographical memories. Cortex 39:643–665
Metzler-Baddeley C (2007) A review of cognitive impairments in dementia with Lewy bodies relative to Alzheimer's disease and Parkinson's disease with dementia. Cortex 43:583–600
Mimura M, Kinsbourne M, O'Connor M (2000) Time estimation by patients with frontal lesions and by Korsakoff patients. J Int Neuropsychol Soc 6:517–528
Miotto EC, Savage CR, Evans JJ, Wilson BA, Martins MG, Iaki S, Amaro EJ (2006) Bilateral activation of the prefrontal cortex after strategic semantic cognitive training. Hum Brain Mapp 27:288–295
Moscovitch M, Rosenbaum RS, Gilboa A, Addis DR, Westmacott R, Grady C, McAndrews MP, Levine B, Black S, Winocur G, Nadel L (2005) Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. J Anat 207:35–66
Northoff G, Heinzel A, de Greck M, Bermpohl F, Dobrowolny H, Panksepp J (2006) Self-referential processing in our brain–a meta-analysis of imaging studies on the self. Neuroimage 31:440–457
Nyberg L, McIntosh AR, Cabeza R, Habib R, Houle S, Tulving E (1996) General and specific brain regions involved in encoding and retrieval of events: what, where, and when. Proc Natl Acad Sci USA 93:11280–11285
Papez JW (1937) A proposed mechanism of emotion. Arch Neurol Psychiatry 38:725–743
Phelps ME (2006) Emotion and cognition: insights from studies of the human amygdala. Annu Rev Psychol 57:27–53
Phelps ME, LeDoux JE (2005) Contributions of the amygdala to emotion processing: from animal models to human behavior. Neuron 48:175–187
Piefke M, Fink GR (2005) Recollections of one's own past: the effects of aging and gender on the neural mechanisms of episodic autobiographical memory. Anat Embryol 210:497–512

Piefke M, Weiss PH, Markowitsch HJ, Fink GR (2005) Gender differences in the functional neuroanatomy of emotional episodic autobiographical memory. Hum Brain Mapp 24:313–324
Piefke M, Weiss PH, Zilles K, Markowitsch HJ, Fink GR (2003) Differential remoteness and emotional tone modulate the neural correlates of autobiographical memory. Brain 126:650–668
Premack D (2007) Human and animal cognition: continuity and discontinuity. Proc Natl Acad Sci USA 104:13861–13867
Rand-Giovannetti E, Chua EF, Driscoll AE, Schacter DL, Albert MS, Sperling RA (2006) Hippocampal and neocortical activation during repetitive encoding in older persons. Neurobiol Aging 27:173–182
Ranganath C, Johnson MK, D'Esposito M (2003) Prefrontal activity associated with working memory and episodic long-term memory. Neuropsychologia 41:378–389
Rugg MD, Otten LJ, Henson RNA (2002) The neural basis of episodic memory: evidence from functional neuroimaging. Philos Trans R Soc B-Biol Sci 357:1097–1110
Schacter DL, Addis DR, Buckner RL (2007) Remembering the past to imagine the future: the prospective brain. Nature Rev Neurosci 8:657–661
Schmitz TW, Johnson SC (2006) Self-appraisal decisions evoke dissociated dorsal-ventral aMPFC networks. Neuroimage 30:1050–1058
Schmitz TW, Kawahara-Baccus TN, Johnson SC (2004) Metacognitive evaluation, self-relevance, and the right prefrontal cortex. Neuroimage 22:941–947
Scoville WB, Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. J Neurol Neurosurg Psychiatry 20:11–21
Skov-Rackette SI, Miller NY, Shettleworth SJ (2006) What-where-when memory in pigeons. J Exp Psychol Anim Behav Process 32:345–358
Staresina BP, Davachi L (2006) Differential encoding mechanisms for subsequent associative recognition and free recall. J Neurosci 26:9162–9172
Steinvorth S, Levine B, Corkin S (2005) Medial temporal lobe structures are needed to re-experience remote autobiographical memories: evidence from H.M. and W.R. Neuropsychologia 43:479–496
Suddendorf T, Corballis MC (2007) The evolution of foresight: What is mental time travel, and is it unique to humans? Behav Brain Sci 30:299–313
Summerfield C, Greene M, Wager T, Egner T, Hirsch J, Mangels J (2006) Neocortical connectivity during episodic memory formation. PLoS Biol 4:e128
Svoboda E, McKinnon MC, Levine B (2006) The functional neuroanatomy of autobiographical memory: a meta-analysis. Neuropsychologia 44:2189–2208
Tulving E (1972) Episodic and semantic memory. In: Tulving E, Donaldson W (eds) Organization of memory. Academic Press, New York, pp 381–403
Tulving E (1995) Organization of memory: Quo vadis? In: Gazzaniga MS (ed) The cognitive Neurosciences. MIT Press, Cambridge, pp 839–847
Tulving E (2002) Episodic memory: from mind to brain. Annu Rev Psychol 53:1–25
Tulving E (2005) Episodic memory and autonoesis: uniquely human? In: Terrace H, Metcalfe J (eds) The missing link in cognition: evolution of self-knowing consciousness. Oxford University Press, New York, pp 3–56
Tulving E, Kim A (2007) The medium and the message of mental time travel. Behav Brain Sci 30:334–335
Uncapher MR, Rugg MD (2005a) Effects of divided attention on fMRI correlates of memory encoding. J Cogn Neurosci 17:1923–1935
Uncapher MR, Rugg MD (2005b) Encoding and the durability of episodic memory: a functional magnetic resonance imaging study. J Neurosci 25:7260–7267
Vandekerckhove MMP, Markowitsch HJ, Mertens M, Woermann FG (2005) Bi-hemispheric engagement in the retrieval of autobiographical episodes. Behav Neurol 16:203–210
Velanova K, Jacoby LL, Wheeler ME, McAvoy MP, Petersen SE, Buckner RL (2003) Functional-anatomic correlates of sustained and transient processing components engaged during controlled retrieval. J Neurosci 23:8460–8470
Viard A, Piolino P, Desgranges B, Chételat G, Lebreton K, Landeau B, Young A, De La Sayette V, Eustache F (2007) Hippocampal activation for autobiographical memories over the entire lifetime in healthy aged subjects: an fMRI study. Cereb Cortex 17:2453–2467

Perception and the Brain

Nikos Logothetis

Abstract There exist numerous explanations for the phenomenon of multistable perceptions (e.g., ambiguous figures or binocular rivalry). Some of the explanations see the answer very early in the visual system as a competition between the monocular retinal inputs. Others like Helmholtz or James, for example, considered attentional mechanisms on higher cognitive levels to be relevant for these phenomena. This article, which is based on a talk presented at the Parmenides faculty meeting 2007, describes and summarizes the main results obtained by electrode measurements of single-cell and open-field activities in different areas of the visual system starting from V1 and V2 in the striate and early extrastriate cortex over V4 and MT up to the inferior temporal cortex. We compare single-cell activities with the reports of the mental perceptions of trained monkeys. The correlations between cell activity and perception increase significantly towards the higher cognitive areas, but are already present within the striate cortex. Our findings suggest that there is no single mechanism for the suppression of visual input but that a series of processes of neural mechanisms at different levels of the visual hierarchy contribute to the overall effect. Even though the article does not address the issue of thinking explicitly, a deeper understanding of how perception is processed in the brain and, in particular, how the correlates of certain neural activities get into the focus of attention and become conscious seems to me a necessary prerequisite for understanding thinking.

N. Logothetis
Dept. Logothetis, Max Planck Institute for Biological Cybernetics,
Spemannstraße 38, 72076, Tübingen, Germany
e-mail: nikos.logothetis@tuebingen.mpg.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_11,

1 Introduction

The study of perception has been a major part of my research work, and in this review I will mainly talk about perception. I will not extend this into thinking and the neural basis of thinking, because the phenomena related to perception are often confusing enough.

I will show that by doing neurophysiological experiments, and subsequently imaging experiments, one can find some very interesting correlations between neural activities of some sort – there are different types of neural activities – and the perceptual phenomena we want to investigate. However, I will also raise the question as to whether this is really bringing us even a single step forward. For the last years, I have been more and more convinced that these correlations may lead us into a dead end and that what we need are better theories and not necessarily more data.

In this article I will mainly report about results of single-cell recordings of neural activities (spike potentials) during the performance of binocular rivalry tasks. These measurements have been made in different visual areas of the cortex of monkeys. One of the conclusions that ensues from this is that there is no such thing as a single neuron or even a single area in the brain which is responsible for the interesting suppression effects in the context of binocular rivalry. We will not make significant progress unless we have a better understanding of the activities of neural networks and their interaction.*

2 Binocular Rivalry Experiments

We know many cases where a physical stimulus is far beyond the threshold and could be perceived very well, and yet it permanently leaves and re-enters our perception. In order to get a better understanding of what is happening in different visual areas, we conducted animal experiments (Logothetis and Schall 1989; Leopold and Logothetis 1996; Sheinberg and Logothetis 1997). Furthermore, we used the paradigm of binocular rivalry which, in general, allows a better control of the experimental situation. In this case, two patterns that are obviously very different were presented (Fig. 1). For very interesting physiological reasons these patterns never fuse into some kind of transparency. It seems that already very early in the visual system there are certain assumptions about the world that are literally instantiated in the connectivity and the interactions in the visual areas, and because logically one cannot have two objects at the same time in one place, the system does not permit any kind of perception of transparency here. We perceive the two images in alternation.

During the last years we investigated the problem of what happens in the brain when the perception of a stimulus is suppressed. Does it entirely disappear, leaving no representation in the visual areas? Or are there active representations of that

* For further details as well as a comprehensive list of references see Blake and Logothetis (2002), Leopold and Logothetis (1999), Leopold et al. (2005)

Fig. 1 Binocular rivalry. In a simultaneous presentation of two different types of stimuli, only one of them is perceived consciously while the other is suppressed. After a few seconds the conscious perception switches and the other image is suppressed

particular perception, which do not reach the level of activity that corresponds to conscious perception?

In many experiments, we did not use the free running rivalry which I just described but we presented one of the objects to one of the eyes first (the so-called dichoptic presentation), and 500–600 ms later we presented the other object to the other eye. The advantage of this dichoptic set-up is that, after the onset of the second object, one is guaranteed to perceive this object. The so-called flash suppression of the first object is aligned with the onset of the second stimulus. This is much more convenient for studying the changes in neural activity (Sheinberg and Logothetis 1997; Wolfe 1984). As far as the physiological mechanisms are concerned, the two situations (the free running rivalry and the rivalry with triggered flash suppression) are interchangeable.

The kind of perception in binocular rivalry is similar (but not identical) to the satiation experiments in the context of bistable perception that psychologists did for many years. An example is shown in Fig. 2. In the ambiguous version one sees a circle (that is perceived as an ellipse) together with a bar. In roughly half of the cases one perceives the bar as going from left to right into the circle and in the other half the bar seems to be going from right to left. If, however, one is primed for a few seconds with the unambiguous version of the image first (Fig. 2b), and then one is presented the ambiguous version, one sees the bar going into the opposite direction.

These are some of the phenomena for which we studied the physiological activities. For our experiments we used trained monkeys. I will not describe how these monkeys are trained (see e.g., Logothetis 1999), but you can be absolutely sure that the monkeys are indeed reporting what a human would report. To train these animals appropriately is a whole science in its own right. We are convinced that when these monkeys report perceptual changes, these are really perceptual changes,

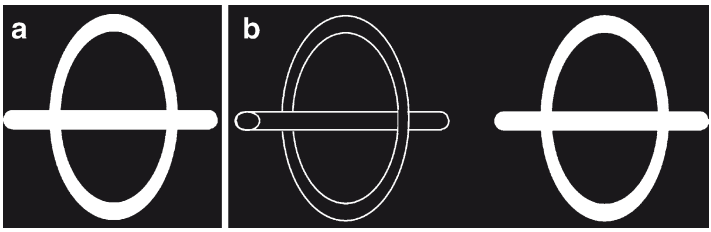

Fig. 2 Ambiguous perception. In the ambiguous figure (**a**), we sometimes perceive the bar as going from right to left through the circle and sometimes the other way round. If we are primed for a few seconds with the unambiguous version (**b**), we will perceive the bar going into the opposite direction when the ambiguous version is presented

just like those experienced by human observes. They do not randomly hit the lever. This is clear from the temporal characteristics of the perceptual alternations as well as from the almost identical psychometric functions of humans and monkeys in experiments, in which stimulus strength is varied and its effect on alternation rate is examined.

Naturally correct behavior follows extensive training and the utilization of many different clever tricks that can be used as telltale signs of the animal's perception. Using all these behavioral tricks, we trained the animals to pull and hold one lever when they see one pattern and to pull and hold the other lever when they see the other pattern. We also trained them to refrain from holding or pushing levers if they see a mixture of the two, which happens for 200–500 ms during the transition time, because we wanted to be as sensible as possible to the changes in the cell activity.

3 Extracellular Recording in the Visual Cortex

We recorded from the visual cortex, which for monkeys is almost identical to the human cortex (Fig. 3, *left*). The information starts in the retina. Here one has a beautiful map of the stimulus, and it proceeds to the geniculate body where one finds another map which is isomorphic to the first one in a precise mathematical sense. Then there is the optical radiation that goes back to the visual area V1. The main difference in the visual cortex between humans and monkeys is that the monkey fovea is projected more to the front compared to humans. The growth of the frontal lobe during the evolution of the human brain pushed some of the visual association in the primary visual cortex backwards. These relations can also be shown nicely in MRI-images (Fig. 3, *right*).

However, the primary visual cortex, or V1, is just the beginning of the story. In Fig. 4a, one recognizes different colors and different names which correspond to the different visual areas. In humans and monkeys about 40–45% of the neocortex is visual. Obviously, a lot of cortical machinery is devoted just to vision. By digitizing the brain and unfolding the image one can generate maps like the one in Fig. 4b

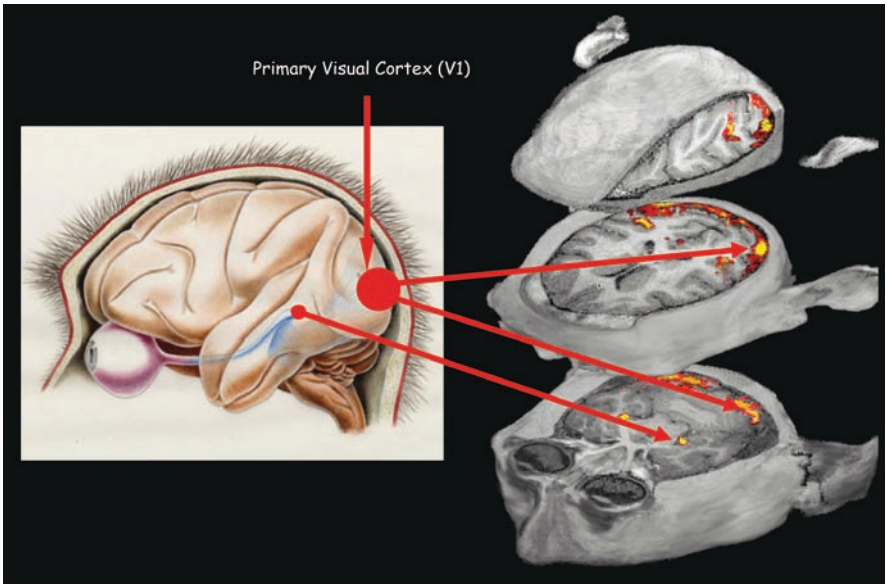

Fig. 3 (*left*) The optical path in the brain of a monkey. From the retina the optical nerves go to the left and right geniculate nuclei. The optical radiation goes back to V1. In the human cortex, the fovea projects further back due to the growth of the frontal lobe. (*right*) MRI-image of the brain of a monkey. The two geniculate bodies of the thalamus light up as well as the areas of the primary visual cortex (V1)

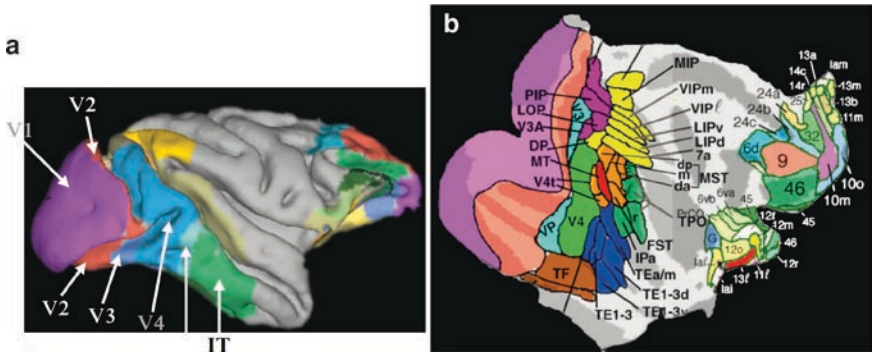

Fig. 4 (**a**) The different areas in the visual cortex. (**b**) By digitizing and unfolding this image one obtains very precise maps of the different visual areas

which shows the positions of the different areas (Lewis and van Essen 2000). It is possible to tell exactly from which of these areas one is recording.

There are many recordings from the different visual areas during the execution of the tasks I mentioned in the beginning. It is well known that a lot of cells fire even if the subject is under anesthesia or sleeping. We have long known that we are

mostly unaware of the activity in the brain that maintains the body in a stable state – one of its evolutionarily most ancient tasks. Our experiments show we are also unaware of much of the neural activity that generates our conscious experiences. The surprising result from electrophysiology is that there are many neurons that continue to be stimulus-selective in conditions in which we have no conscious experiences. The following questions are thus reasonable: Are there active neurons, that determine whether you see an object or not? Are the active neurons everywhere or are they concentrated in one area? Is this area controlling everything?

Our laboratory mostly made extracellular recordings. This is important, because it is essentially the methodology that I will put under criticism.

When a stimulus – basically a change in the concentration of neurotransmitters – enters a neuron it causes a depolarization of the membrane. An electrode outside but close to the neuron senses a negativity, because the positive ions flow into the neuron. Due to Poisson's law, these currents run around in loops and, therefore, there will be a positivity far away in the non-activated area. These loop currents generate dipole fields which lead to a voltage difference in the order of a few hundred microvolts that can be measured very precisely. If there were only one neuron, we could deduce the cell activity quite well by measuring these voltage differences. However, most of us happen to have more than one neuron which makes things very complicated because different dipole fields can influence and even annihilate each other.

Basically, the fields depend on the geometry of the cell arrangement. However, the geometry of the cortex, as well as the geometry of the hippocampus, the cerebellum and other areas, helps us very much. As one can see in Fig. 5 (*left*), the fascicles of the apical dendrites run together from deeper layers to upper layers in very strong bundles (Peters and Kathleen 1994). These bundles generate a so-called open field where one charge is in one location and the other type of charge is in another location (Fig. 5, *right*). This can be measured and with the appropriate electrode one finds very strong sum potentials (Logothetis et al. 2007).

If the electrode is very close to a neuron one can actually measure something that is called extracellular spike- or extracellular action potential. A simple mathematical transformation of this spike potential allows the precise determination of the intracellular action potential that has been measured long ago by biophysicists. This kind of recording is what people, including us, have been doing for many years, basically ignoring all other information – unfortunately.

The electric signal is characterized by time-varying spatial distributions of action potentials superimposed on relatively slow varying field potentials. (Fig. 6a). This signal is turned into a binary function by detecting the spikes and setting the function to 1 at their time of occurrence; otherwise to zero (Fig. 6b). One then generates a representation of the instantaneous rate by counting the number of spikes produced within a time bin of the order of 20, 50 or 100 ms (Fig. 6c).

Figure 7 shows a recording from the area V1 in the striate cortex together with the reporting of the monkey (Leopold and Logothetis 1996). The data in Fig. 7a correspond to an unambiguous presentation of different stimuli. As one would expect, the tuned cells fire if the stimulus corresponds to the preferred orientation

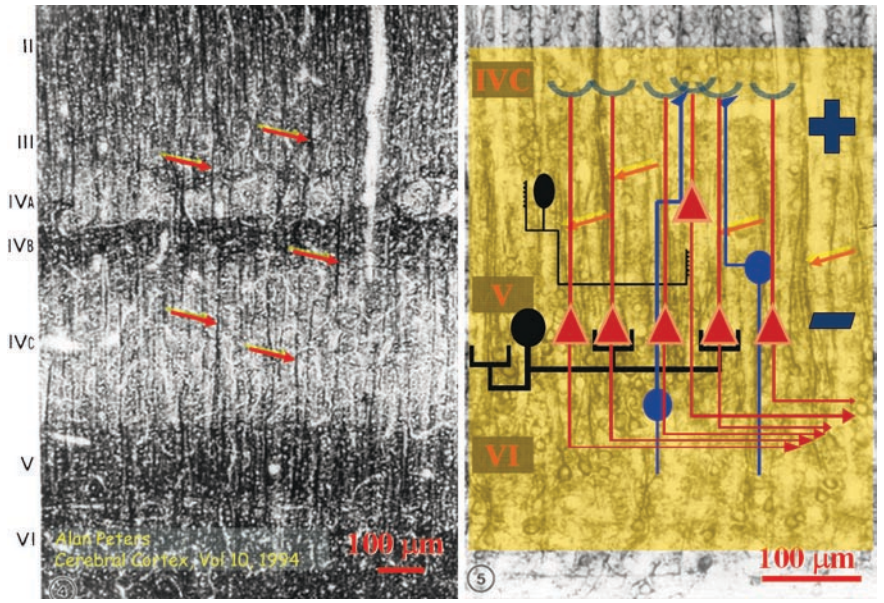

Fig. 5 Bundles of fossicles of apical dendrites generate the so-called open field

of the cell and they do not fire in the other case. Figure 7b shows the data when the stimulus is presented in a binocular rivalry task. Now the firing of the cells is quite random and independent of the reporting of the animal.

The situation changes quite drastically if one records from the areas V4 or MT, where V4 represents mostly the color and form of objects, and MT represents information about motion. In this case, the neurons tend to fire about 300–400 ms before the animal reports the preferred orientation. This corresponds roughly to the latency of the animal. So, the firing rate of the neuron is correlated to the bars which indicate what the animal sees (Fig. 8).

While in the striate cortex about 10–13% of the neurons are active during a stimulation, in the extrastriate cortex about 40–45% of the neurons participate (Leopold and Logothetis 1999). They modulate their activity according to the perceptual changes. But about half of them will explicitly fire only when the preferred stimulus is hidden. These results show that there is a representation of both stimuli in the extrastriate cortex. However, only one of them reaches consciousness and the other doesn't and remains hidden.

One step higher one arrives at the so-called inferior temporal cortex, where people, including ourselves, have described the existence of very complex physiological properties. In this area, cells will only fire if the stimulus corresponds to complex objects – faces, hands, etc. In the recordings, we find cells whose activity is correlated with the report of the monkey (Sheinberg and Logothetis 1997). During the rivalry task (gray area in Fig. 9) one observes a long period where the cell is not firing and then, suddenly, there is an increase

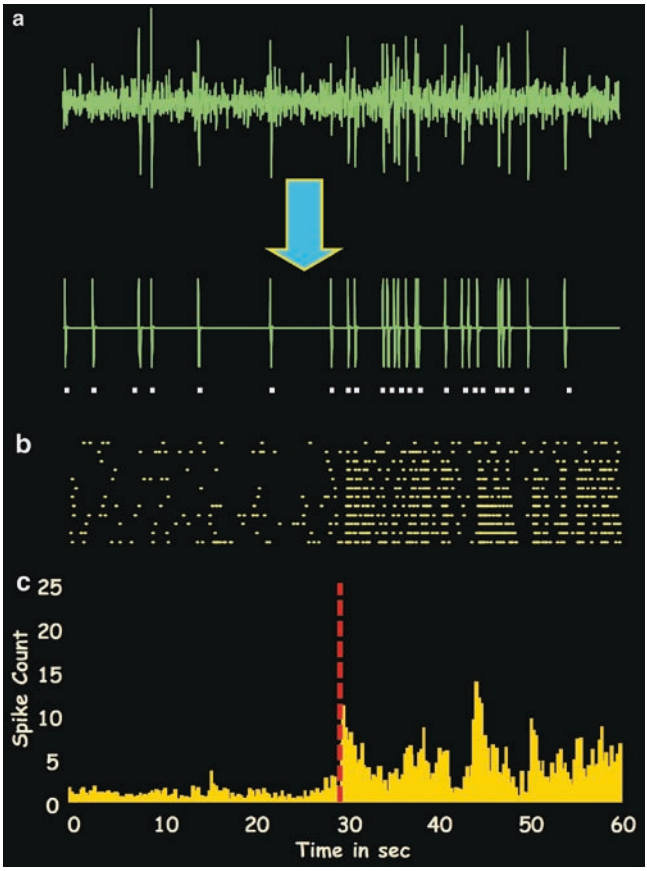

Fig. 6 (**a**) Recorded voltage variation as a function of time, (**b**) binary spike function, (**c**) spike rate

of the cell activity and immediately after this increase the report of the monkey comes and indicates: "I see the right stimulus," which happens to be the preferred stimulus of the cell.

In the inferior temporal cortex over 95% of the neurons participate in this activity. The neurons will fire exclusively when the preferred stimulus is consciously perceived and they will not fire for the hidden stimulus. All the mutually antagonistic interactions – excitatory and inhibitory activations or membrane potentials – happen in the early visual areas, but not at all in the very late visual areas. The inferior temporal cortex is the last station where exclusively visual information is processed in the cortex. After that, all other areas are multimodal.

What do we learn from all that? The most important lesson is that no single area alone is responsible for the suppression. There was a long-lasting psychophysical discussion claiming that there is an area where this suppression occurs (for a history of this discussion see e.g., Blake and Logothetis 2002). It seems to me that, at least

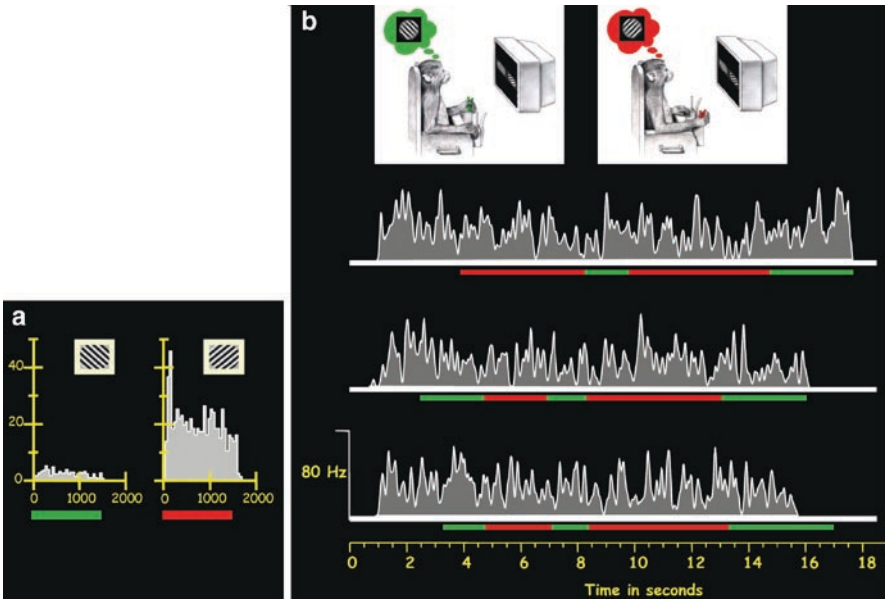

Fig. 7 Recording from the area V1. The bars underneath the instantaneous firing rate of the neuron show the periods where the animal reports seeing the prefered orientation of the cell. (**a**) Firing rate during the unambiguous presentation of the stimulus, (**b**) firing rate during the binocular rivalry task

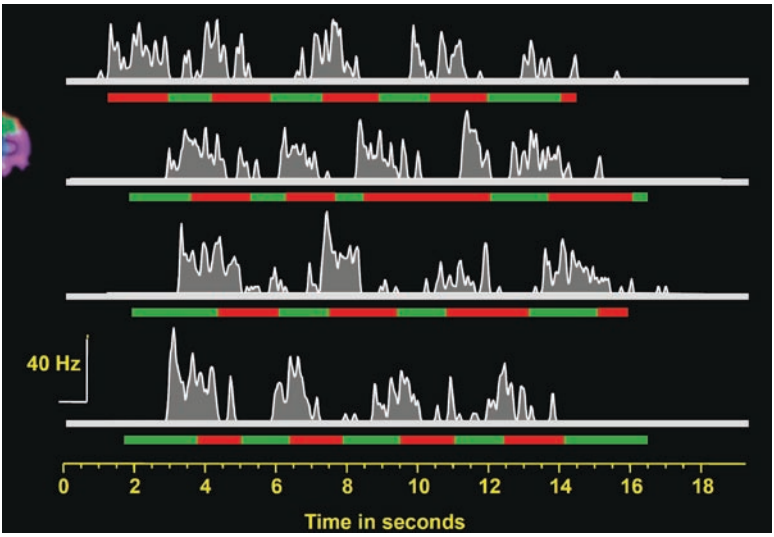

Fig. 8 Recording from the visual areas V4 and V5. The activity of the neurons is correlated to the reports of the monkey. Some of the neurons explicitly fire only when, according to the animals report, the preferred stimulus is hidden

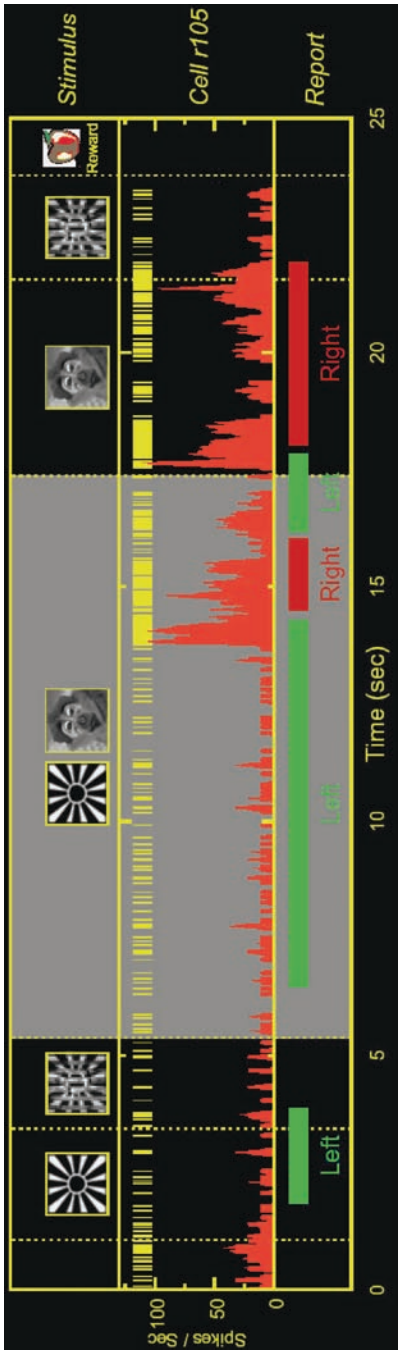

Fig. 9 Recording from the inferior temporal cortex. The gray area corresponds to the rivalry situation; in the period before and after the stimulus is unambiguous

from a physiological point of view, we cannot hold this claim, because the activity is distributed all over the place.

From these results we come to the following conclusions:

- The vast majority of V1 neurons are active whether or not the stimulus is perceived. That explains a huge number of psychophysical effects which will not be discussed here.
- The cells in IT follow the sequences of perceptual dominance and suppression.
- The neurons in the early extrastriate cortex may fire selectively for the dominant or for the suppressed stimulus.
- In primary and early extrastriate cortices, small changes in the firing of cells may be sufficient to instigate a perceptual shift.
- The interneuron response-coherence is related to the system's stability.

Concerning the last point one can show that the synchronization between cells breaks down in the ambiguous situation, and there is considerable synchronization even in the spontaneous activity for an unambiguous stimulation.

4 Rivalry Related Tasks

As far as the physiology is concerned, all these results seem to be very general, whether one considers binocular rivalry, bistable perception like the Necker cube (Necker 1832), or figure-ground illusions like in the case of the vase and the face (Rubin 1958). I will describe two more examples related to the perception of objects.

The set-up of the first example is again a dichoptic presentation where one eye looks at some object (e.g., a red circle) and the other eye looks at a background of moving dots (Fig. 10), which do not intersect with the area where the object is presented (Wilke et al. 2003). Furthermore, object and background are not presented simultaneously, but the object is presented first and with a little time shift the background is flashed on. During the flashing of the background, the red stimulus sometimes disappears from our conscious perception. In this case, there is no rivalry, because there is no coexistence of forms (the dots of the background do not enter the part of the visual perception which is occupied by the object in the other eye).

It turns out that the probability of disappearance of the red object, the so-called mask effect, depends on several parameters: it depends on the density of the moving dots, it depends on the distance the moving dots keep from the central object (the larger this distance, the later the onset of the disappearance and the later the onset of the neural responses), and, finally, it depends on the target-surround asynchrony. It does not depend on the nature of the object; what is going to disappear could be virtually anything: a face, an object, several objects, etc. The probability of disappearance is maximal (almost a hundred percent), if there is a shift of one second between the onset of the first image – the object – and the flashing of the background.

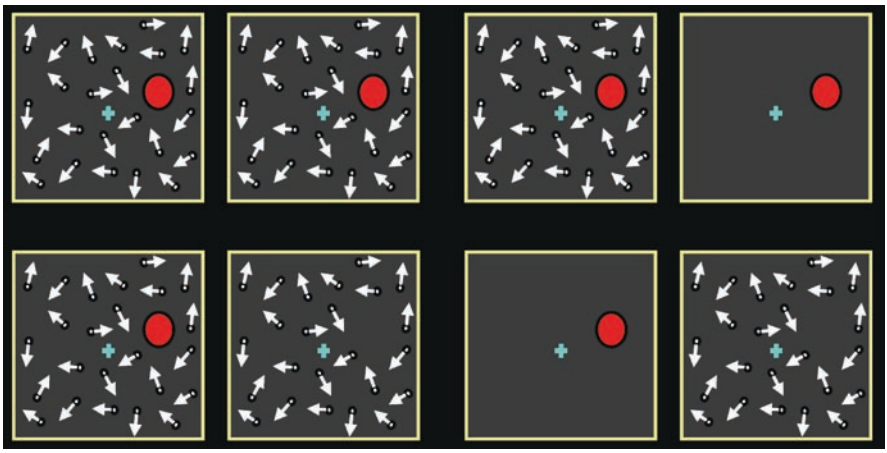

Fig. 10 In a dichoptic asynchronous presentation, the perception of the object (*left*) sometimes vanishes with the onset of the background (*right*)

The second example also exhibits interesting relationships between cell activation in the inferior temporal cortex and actually seeing the object. We are all familiar with the situation of searching for our keys: We look a hundred times and probably even gaze at the keys during this search but we do not see them. Then at some point there is an "Aha!" and we find the keys.

This is exactly what the monkey does in the following task. The animal is supposed to find the cat in the picture of Fig. 11; it knows the object. In general, monkeys are very good at detecting embedded objects. In the beginning, the gaze of the monkey has a distance of ten degrees from the embedded object. The monkey looks around and at some stage the distance becomes very small – the object is within the fovia – but there is no cell activity and there is no response of the animal. Then, a few hundred milliseconds after the gaze of the monkey has left the site of the object, there is a sudden increase of cell activity and immediately after that the monkey gives the response that he sees the object. So, even if the gaze of the monkey is right on the object, the cell doesn't fire – not until the monkey sees the object consciously.

5 Things We Learn and Things We Do Not Learn

Once more I raise the question: What do we learn from these data? This is a problem that has captured me for the last 7–8 years. There is no question that we learn a lot about cell properties. We now have a huge repertoire of results and with these results we can basically characterize cells, and we even get some ideas about the complexity of the connectivity of the cell. Furthermore, we learn about areas that may or may not be related to particular tasks or to particular stimuli.

Fig. 11 Hidden in the picture is the image of a cat which the monkey is supposed to find. Below one sees the distance of the gaze of the monkey (measured as an angle) from the object. At one point, the gaze is right on top of the object, but only after the cell activity increases does the monkey report seeing the object

We learn something about the percentage of neurons showing certain characteristics. We know that particular areas have a lot of neurons that respond to the color red and other areas have a lot of neurons that respond to movement. Furthermore, we learn something about the effects of mental states on the firing neurons. People have described effects of attention, short-term memory, and all kinds of things that seem to have an impact on the cell.

The following is a list of things which we do not learn, however:

- *Computational rules.* We have almost no idea how the things I mentioned actually do happen. On the other hand, we will never really make progress unless we have a better understanding of this. What we do today is some kind of microphrenology: describing correlates of everything one can imagine.
- *The role of recurrence, feedback, excitation – inhibition.* We often pretend that the cortex is just there, and when an input is coming in, there is some data processing in the cortex and it sends the results somewhere else. But the input, usually from the previous site, the thalamus or another area, is extremely small. Then there is an amplification with a positive feedback. It could lead to a run-away excitation, but it is very tightly controlled by inhibitory neurons. Only the deviations in this balance between excitation and inhibition create a response.

To understand these mechanisms is much more difficult than to simply assume that the neuron is just sitting there and is doing absolutely nothing until the stimulus comes in and then it gets excited.

- *The role of interneurons.* Interneurons are very crucial in all these processes but I will not go into detail about this.
- *The cell-network-behavior.* This is a very interesting relationship whose relevance is only slowly accepted by the community. At present, we often try to get rid of the "network." We assume that there is only the cell and the behavior. But this assumption is ridiculous. Instead, we should think of the cell as having a crucial role within the network, which is very interesting to understand, and then the network plays a role in the behavior.

To make the last point clearer, let me describe what happens when a stimulus is presented. For instance, we presented objects with different cues, and we tried to see whether there is some kind of cue constancy or shape dependence. One observes a widespread activation which is far from being random: in repeated presentations of the same stimulus one always finds the same activated spots. One also observes local activations, but at the same time one finds these activations in many different areas.

We also know that these areas are incredibly interconnected, sometimes in the most confusing way. One area separates the input and sends the information to the next area, and then the next area respects nothing about this separation of the first area. Obviously the brain does the job, but not in the simple way that we sometimes want it to.

Furthermore, the cortical manifold is interacting a lot with different substructures. We know that, for example, pulvinar may play a critical role in synchronizing different cortical areas that are supposed to respond together, and the Claustrum also plays a very important role in the communication between these areas. We have only just started to understand some of the relevant interplay between the different parts of the brain.

6 Conclusion

I have used the example of binocular rivalry and dichoptic presentations to illustrate the complicated interplays between different areas of the brain which lead to the difference between what we see and what we have in mind. Not only is the step from neurons to perception a huge one, but there are all kinds of intermediate levels which we are just about to begin to understand.

New methods which we are presently developing may help to collect more data about these intermediate levels. Some of these methods include intracortical recording, physiological and functional cerebral blood flow studies, studies using neuromodulator injections, the investigation of the functional neurovascular coupling, microstimulation and fMRI, perfusion and hypercapnia, spectroscopic imaging, the investigation of structural neurovascular coupling, investigation of the connectivity

with MR Tracers, and, finally, molecular imaging, i.e., the development of chemical substances that can substitute hemodynamic responses with either calcium or voltage images. But, as I said in the beginning, in the end what we need are not necessarily more data but a theory and a plausible theoretical context within which data can be better (and more intelligently) interpreted.

References

Blake R, Logothetis NK (2002) Visual competition. Nat Rev Neurosci 3:1–11

Carmichael ST, Price JL (1994) Architectonic subdivision of the orbital and medial prefrontal cortex of the macaque monkey. J Comp Neurol 346(3):366–402

Leopold DA, Logothetis NK (1996) Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. Nature 379:549–553

Leopold DA, Logothetis NK (1999) Multistable phenomena: changing views in perception. Trends Cogn Sci 3:254–264

Leopold DA, Maier A, Wilke M, Logothetis NK (2005) Binocular rivalry and the illusion of monocular vision. In: Maass W, Bishop CM (eds) Pulsed neural networks. MIT Press, Cambridge, MA

Lewis JW, van Essen DC (2000) Corticocortial connections of visual, sensimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. J Comp Neurol 428:37–112

Logothetis NK (1999) Vision: a window on consciousness. Sci Am 281:68–75

Logothetis NK, Schall JD (1989) Neuronal correlates of subjective visual perception. Science 245:761–763

Logothetis NK, Kayser C, Oeltermann A (2007) In vivo measurement of cortical impedance spectrum in monkeys: implications for signal propagation. Neuron 55:809–823

Necker LA (1832) Observations on some remarkable optical phaenomenon seen in Switzerland; and on an optical phaenomenon which occurs on viewing a figure of a crystal or geometrical solid. Lond Edinb Philos Mag J Sci 1:329–337

Peters A, Rockland KS (Eds.) (1994) Primary visual cortex in primates (Vol 10). Cerebral Cortex. New York: Plenum

Rubin E (1958) Figure and ground. In: Beardslee DC, Wertheimer M (eds) Readings in perception. Princeton, NJ, Van Nostrand

Sheinberg DL, Logothetis NK (1997) The role of temporal cortical areas in perceptual organization. Proc Natl Acad Sci USA 94:3408–3413

Wilke M, Logothetis NK, Leopold DA (2003) Generalized flash suppression of salient visual targets. Neuron 39:1043–1052

Wolfe DL (1984) Reversing ocular dominance and suppression in a single flash. Vis Res 24:471–478

Part III
Onto- and Phylogenetic Considerations

A Developmental Perspective on Modularity

Annette Karmiloff-Smith

Abstract The notion that the mind/brain is composed of independently functioning modules may hold to some extent for the adult brain, once it has become fully specialised or if it displays acquired domain-specific deficits when focal damage has occurred. The extension of this thinking to typically and atypically developing infants in terms of innately specified, intact or impaired modules is not, however, warranted. This chapter discusses modularity from a developmental perspective and shows how specialisation and localisation of cognitive and brain function occurs very progressively over ontogenetic time. In other words, it argues for a gradual process of *modularisation*, not built-in modules.

1 Introduction

The idea that the brain is composed of specialised, independently functioning modules has a long history. Dating back to Gall's phrenology, it became the central thesis of the "boxology" model of acquired adult neuropsychology in which the brain's functioning is represented by a series of boxes and arrows, with impaired boxes crossed through. As brain-damaged adults with uneven neuropsychological profiles were identified, so researchers divided the mind/brain into separated boxes for number, face processing, space, semantics, syntax and so forth, each processed within a purported specialised region of the brain. But it was really in the early 1980s, with the publication of Fodor's "Modularity of Mind" (Fodor 1983), that the modular notion fully permeated both adult and developmental psychology as well as cognitive neuroscience.

A. Karmiloff-Smith
Centre for Brain and Cognitive Development, Birkbeck, University of London, 32 Torrington Square, London, WC1E7HX, UK
e-mail: a.karmiloff-smith@bbk.ac.uk

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_12,

Fodor invoked nine criteria, all of which had to be met for something to qualify as a "module":

1. *Domain specificity:* a given brain module can only process proprietary inputs from a specialised domain
2. *Mandatory processing:* the brain cannot control modular processing in a voluntary fashion
3. *Limited central access to intermediate representations:* the brain has no conscious access to the inner workings of a module
4. *Speed:* modular processing is extremely rapid, compared to conscious thought
5. *Shallow output:* the outputs yielded by modular processing are low-level and only of use within the narrow confines of a particular module
6. *Fixed mental architecture:* modular processing is carried out by a dedicated brain region
7. *Patterns of characteristic breakdown:* in focal brain damage, there is no overall loss of capacities, but selective modular deficits
8. *Characteristic pace and sequencing in development:* each module is innately specified, develops independently, via maturation of specific brain regions, and is insensitive to environmental influences
9. *Information encapsulation:* information being processed within a module cannot be accessed by another system in the brain.

It is important to note that Fodor applied the notion of modularity to perceptual input systems, including language (Fodor 1983), whereas subsequent authors have extended the modular concept to higher cognitive-level abilities as well as to output systems. Moreover, while some authors have subsequently relaxed a number of the above criteria in order to enable the generalisation of modular thinking to cognitive-level processes (Sperber 2001; Leslie 1992), for Fodor a perceptual input module had to meet every one of the nine criteria.

2 Modularity, Evolution and Development

It was particularly researchers working on one version of so-called Evolutionary Psychology who honed in on Fodor's modularity concept to claim that much of our ancestral past could be understood in terms of modules passed on through evolution, as is the case, for instance, of a "cheater module" (Duchaine et al. 2001). For these authors, the brain was conceptualised in terms of the metaphor of a Swiss army knife, each tool being exquisitely fashioned and dedicated to carrying out a very circumscribed task, passed on by Evolution from our hunter-gatherer ancestors.

But one of the most ardent uses of the modularity concept came from those studying acquired deficits and developmental disorders of genetic origin (e.g. Baron-Cohen 1998; Baron-Cohen et al. 1986; Frith 1986; Temple 1997). Because of the similar patterns of deficits and proficiencies of some developmental disorders and those seen in adult neuropsychological patients, the concept of a modular mind/brain tended to take over the field of developmental cognitive neuroscience

in the final decade of the twentieth century (e.g. Baron-Cohen 1998; Baron-Cohen et al. 1986; Frith 1986; Temple 1997). General intelligence tests became less favoured than those which could be interpreted in modular terms. So, on a given test of a specific domain, behavioural scores that fell "in the normal range" were considered to involve an "intact module", whereas those that fell below the normal range were explained by the notion of an "impaired module". Acquired deficits were conceptualised as damage to a specific module in the brain, whereas genetic disorders were considered as mutations to the genes that purportedly built the specific modules.

Only 3 years after Fodor's 1983 publication, a new, modular explanation of autism was offered. Baron-Cohen, Leslie and their colleagues claimed that autism could be explained by the lack of, or damage to, a theory-of-mind module (Baron-Cohen et al. 1986; Leslie 1992), impaired by a specific set of mutated genes which interfered with the development of a specific region, the orbito-frontal cortex, claimed to be involved in computations for the attribution of intentional states to others (Baron-Cohen et al. 1999). Rapidly the explanatory concept of damaged versus intact modules was extended to developmental disorders in general (Baron-Cohen 1998; Temple 1997) and to particular disorders with uneven cognitive profiles, such as dyslexia (Frith 1986; Castles and Coltheart 1993), Specific Language Impairment (Gopnik 1997), Williams syndrome (Bellugi et al. 1994), developmental dyscalculia (Butterworth 2005a; Temple 1997), and developmental prosopagnosia (Young and Ellis 1989; Duchaine 2000).

Modular explanations were also extended to studies of typically developing children by researchers of a Nativist persuasion (e.g. Spelke 1998). Any time a competence was detected within the first few months of life, an explanation was not sought in the infant's early capacity for learning. Rather, infants were claimed to be born with an innately specified module for that domain: number (Dehaene 1997; Butterworth 2005b; Gelman 1993), face processing (Duchaine 2006), language (Pinker 1994), spatial cognition (Hermer and Spelke 1996), and knowledge of the constraints governing the physical world (Spelke 2005). Learning, as such, was rapidly banished from having any explanatory role (Piatelli-Palmerini 2001).

Thus, competences found in typically developing infants, impairments identified in brain damaged adults and in children with genetic disorders, as well as arguments from Evolutionary Psychology, all seemed to corroborate the claim that the human mind/brain is composed of highly specialised, independently functioning input and output systems, at both the perceptual and cognitive levels. So what is wrong with that position? A very different view is found in research espousing a Neuroconstructivist position.

3 Gradual Developmental Process of Modularisation

Neuroconstructivism argues that if the adult brain is in any way modular, this is the product of an emergent developmental process of modularisation, not its starting point (Karmiloff-Smith 1992, 1997, 1998, 2007; Elman et al. 1996; Johnson et al. 2002; Mareschal et al. 2007). A crucial error is to conflate the specialised brains of

adults, which have developed normally prior to damage in later life, with those of infants and children, which are still in the process of developing (Karmiloff-Smith et al. 2003). To date there is no evidence to suggest functional specificity of gene expression in the brain, i.e. no evidence that genes which are expressed in the brain solely target discrete cortical regions. Rather, gene expression is widespread showing diffuse, large-scale gradients across cortex (Kingsbury and Finlay 2001; Karmiloff-Smith 2006). Thus, genetic mutations contributing to developmental disorders in infants are likely to affect widespread systems within the brain (Karmiloff-Smith 1998). This does not preclude that the outcome of the dynamic developmental process could result in some areas being more impaired than others, due to the processing demands of certain kinds of inputs to those areas and to differences in synaptogenesis across various cerebral regions (Huttenlocher and Dabholkar 1997). Moreover, the innate modular view tends to underestimate the changing patterns of connectivity within and across different brain areas during development. Indeed, the same behaviour may be subserved by different neural substrates at different ages during development (Karmiloff-Smith 1998).

In studies of typically developing infants and of those with developmental disorders, researchers have shown how different cortical pathways become increasingly specialised and localised as a result of being recruited for specific tasks over developmental time (Elman et al. 1996; Johnson 2001). Various areas of the brain start out by competing to process different inputs (Karmiloff-Smith 1998), because cortical regions initially respond to a wide variety of different stimuli and task situations. In other words, the infant brain displays more widespread activity than the older child or adult brain when processing specific kinds of inputs. With time, however, the developing brain starts to show increasing specialisation and localisation of function as certain areas win out in the competitive processing. How does Neuroconstructivism explain this? Starting out with tiny differences across brain regions in terms of the patterns of connectivity, the balance of neurotransmitters, synaptic density, neuronal type/orientation and the like, some areas of the brain are somewhat more suited (i.e. more relevant in terms of their computational properties) than others to the processing of certain kinds of input, and over time they ultimately win out. These pathways do not start out as domain specific, however. They start out as "domain relevant". In other words, the computational properties e.g. (types of neurons, density of neurons etc.) of a particular brain circuit may be more relevant to certain types of processing (e.g. holistic vs. componential processing) than others, although they are initially not specific to that type of processing only. It is only after developmental time and repeated processing that such a circuit *becomes* domain specific as ongogenesis proceeds (Karmiloff-Smith 1998). There is thus a gradual process of recruitment of particular pathways and structures for specific functions (Elman et al. 1996), such that brain pathways that were previously partially activated in a wide range of task contexts increasingly confine their activation to a narrower range of inputs and situations (Johnson et al. 2002). The next section provides a concrete example of progressive modularisation in the typical case and how it may fail in the atypical case.

This neuroconstructivist approach, examining the interactions between areas and their temporal and spatial dynamics over developmental time, should replace the misguided search for pre-specified modules. Yet the association between the location of brain damage and cognitive-level deficits has been the central approach in traditional cognitive neuropsychology and continues to characterise much of the work on developmental disorders (see critiques in Karmiloff-Smith 1997, 1998; Karmiloff-Smith et al. 2003).

4 A Concrete Example of Progressive Modularisation: Face Processing in Typically and Atypically Developing Populations

Studies of typically developing infants (Johnson and de Haan 2001; de Haan et al. 2002; Cohen-Kadosh and Johnson 2007) show that in early infancy both the left and the right ventral visual pathways are differentially activated by faces. Over a lengthy period of development, face processing in most adults ultimately localises to the right ventral pathway. Likewise, word recognition initially invokes widespread cortical activity, but with development, gradually localises to the left temporal region (Neville et al. 1994). This is not a question of maturational age, but of experience of processing particular types of inputs.

Looking at a specific example, one developmental disorder, Williams syndrome (WS), has been signalled out for its behavioural proficiency in face processing tasks. Several labs across the world have shown that individuals with this disorder, involving a deletion of some 28 genes on one copy of chromosome 7 (Donnai and Karmiloff-Smith 2000), display scores within the normal range on tasks such as the Benton Face Recognition Task and the Rivermead Face Memory Task (Bellugi et al. 1994; Udwin and Yule 1991). Surely, in a clinical population with an average IQ in the mid 50s, this prowess at face processing must be evidence for an intact, innately specified face-processing module that the genetic mutation has not affected? More in-depth behavioural and brain studies of WS show this to be a premature conclusion because it fails to distinguish between behavioural scores and underlying cognitive processes (Karmiloff-Smith et al. 2004). The same behaviour can be achieved by different processes as a result of different cognitive trajectories over developmental time (Annaz et al. 2008; Ansari and Karmiloff-Smith 2002; Cornish et al. 2007; Karmiloff-Smith et al. 2004). Our research showed that, despite achieving "normal" scores, individuals with WS displayed reduced sensitivity to inverted faces compared to the inversion effect that typically developing children show over developmental time. In other words, individuals with WS did not become more proficient at upright faces in comparison to inverted ones, even by the time they reached adulthood (Karmiloff-Smith et al. 2004). Our studies of the electrophysiology of the WS brain also revealed that this clinical group failed to display the progressive right hemisphere localisation when processing faces (Grice et al.

2001, 2003). They also failed to display specialisation of function, i.e. a different pattern of brain activation when processing faces compared to cars, which typical controls did display (Grice et al. 2001, 2003; Mills et al. 2000).

In summary, the proficient face processing of individuals with Williams syndrome cannot be taken as evidence for an intact face processing module, innately prespecified and passed down through evolution. Rather, as with many other cognitive-level functions, face processing is an *emergent function that develops over time*, a developmental process which follows an alternative pathway in this clinical group. Indeed, the gradual process of development must always be taken into account (Karmiloff-Smith 1998, 2007).

5 Concluding Thoughts

Fodor's notion of a module has greatly influenced psychology and cognitive neuroscience. However, it is a static concept that fails to take the process of ontogenetic development into account. Infants are not born with pre-specified modules. Indeed, the infant cognitive system is less differentiated and thus less modular than the adult system, suggesting that modularity is an emergent property of the developmental process. So, domain specificity is not a built-in property of the brain but emerges over developmental time. And even if a modular organisation of the adult brain is the emergent outcome of development, even adult modules should not be viewed in terms of the rigid, static notion of a Fodorian module as outlined in the introduction. Thus, instead of the notion that a given brain module can only process proprietary inputs from a specialised domain, Neuroconstructivism argues that the brain *becomes* very gradually more specialised over developmental time whereby it narrows its response to the types of inputs a given brain circuit may process, after initially processing many different types of inputs. This is also a relative rather than rigid concept. Indeed, brain circuits that have become relatively domain-specific may still attempt to process new inputs from other domains. The so-called specialised face area in the fusiform gyrus, which becomes face sensitive over developmental time (Cohen-Kadosh and Johnson 2007; Johnson and de Haan 2001), has been shown also to be active even in adults for processing other domains of expertise when repeated training is given (Gauthier et al. 2000). So Fodor's notion of proprietary inputs for specific brain areas has been considerably modulated by subsequent research in both adults and children. Nonetheless, it is true that with development, speed of processing increases as the infant brain becomes more specialised. Neuroconstructivism also modifies Fodor's notion of fixed mental architecture to one in which processing *becomes* increasingly localised in certain areas or circuits of the brain. Again, this is not rigidly so, and throughout life there is a great deal of dynamic on-line reconfiguring of brain circuits to process new inputs. Fodor's notion of patterns of characteristic breakdown once again only holds for the adult brain and even then there appear to be considerable individual differences in the mapping between the brain region(s) damaged and the resulting deficits. Sometimes

impairments are relatively general; rarely are they strictly modular, even in the adult. Once we think of the brain as a dynamic system, constantly undergoing synaptic change, this is hardly surprising. And in the developmental case, brain regions do not simply mature, as Fodor's criteria would have us believe. Their maturation is influenced by gene expression, which in turn is influenced by interaction with other genes and with the environment.

In conclusion, developmental disorders need to be examined within a dynamic view of development rather than being divided into independent intact and impaired parts of a static system. Indeed, in some developmental disorders, increasing specialisation and localisation of function may fail to occur, despite proficient behaviour (Karmiloff-Smith 2007). Thus, the status of a module in adults with genetic disorders can only be understood by looking beneath any behaviours that happen to fall "in the normal range" at the underlying cognitive and brain processes. Importantly, to fully comprehend typical and atypical developmental pathways, the process of ontogenetic developmental change must always be a fundamental part of the explanation of human cognition. To reiterate, the brain is not static; it undergoes extensive developmental changes over ontogenesis (Giedd et al. 1999; Johnson 2001). Finally, it remains an open question as to whether the fully formed adult brain is as strictly modular as some theorists would claim, particularly with respect to the cognitive level, since dynamically changing interactivity of cerebral networks seems always to be the case, not only in children but also in adults.

References

Annaz D, Karmiloff-Smith A, Thomas MSC (2008) The importance of tracing developmental trajectories for clinical child neuropsychology. In: Reed J, Warner Rogers J (eds) Child Neuropsychology: Concepts, Theory and Practice

Ansari D, Karmiloff-Smith A (2002) Atypical trajectories of number development. Trends Cognit Sci 6(12):511–516

Baron-Cohen S (1998) Modularity in developmental cognitive neuropsychology: evidence from autism and Gilles de la Tourette syndrome. In: Burack JA, Hodapp RM, Zigler E (eds) Handbook of mental retardation and development. Cambridge University Press, Cambridge, UK, pp 334–348

Baron-Cohen S, Leslie AM, Frith U (1986) Mechanical, behavioural and Intentional understanding of picture stories in autistic children. Br J Dev Psychol 4(2):113–125

Baron-Cohen S, Ring HA, Wheelwright S, Bullmore ET, Brammer MJ, Simmons A, Williams SCR (1999) Social intelligence in the normal and autistic brain: an fMRI study. Eur J Neurosci 11(6):1891–1898

Bellugi U, Wang P, Jernigan TL (1994) Williams syndrome: an unusual neuropsychological profile. In: Broman S, Grafman J (eds) Atypical cognitive deficits in developmental disorders: implications for brain function. Erlbaum, Hillsdale, NJ, pp 23–56

Butterworth B (2005a) Developmental Dyscalculia. In: Campbell JID (ed) Handbook of mathematical cognition. Psychology Press, pp 455–467

Butterworth B (2005b) The development of arithmetical abilities. J Child Psychol Psychiatry 46(1):3–18

Castles A, Coltheart M (1993) Varieties of developmental dyslexia. Cognition 47(2):149–180

Cohen-Kadosh K, Johnson MH (2007) Developing a cortex specialized for face perception. Trends Cogn Sci 11(9):367–369
Cornish K, Scerif G, Karmiloff-Smith A (2007) Tracing syndrome-specific trajectories of attention across the lifespan. Cortex 43(6):672–685
de Haan M, Humphreys K, Johnson MH (2002) Developing a brain specialised for face perception: a converging methods approach. Dev Psychobiol 40:200–212
Dehaene S (1997) The number sense: how the mind creates mathematics. Oxford University Press, USA
Donnai D, Karmiloff-Smith A (2000) Williams syndrome: from genotype through to the cognitive phenotype. Am J Med Genet 97(2):164–171
Duchaine B (2000) Developmental prosopagnosia with normal configural processing. NeuroReport 11(1):79–83
Duchaine B (2006) Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. Cogn Neuropsychol 23(5):714–747
Duchaine B, Cosmides L, Tooby J (2001) Evolutionary psychology and the brain. Curr Opin Neurobiol 11:225–230
Elman JL, Bates E, Johnson MH, Karmiloff-Smith A, Parisi D, Plunkett K (1996) Rethinking innateness: a connectionist perspective on development. MIT, Cambridge, Mass
Fodor J (1983) Modularity of mind. MIT Press/Bradford Books, Cambridge, MA
Frith U (1986) A developmental framework for developmental dyslexia. Ann Dyslexia 36:69–81
Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000) Expertise for cars and birds recruits brain areas involved in face recognition. Nature Neurosci 3(2):191–197
Gelman R (1993) A rational-constructivist account of early learning about numbers and objects. In: Medin D (ed) Learning and motivation, vol 30. Academic Press, New York
Giedd J, Blumenthal J, Jeffries N, Castellanos F, Liu H, Zijdenbos A, Paus T, Evans A, Rapoport J (1999) Brain development during childhood and adolescence: a longitudinal MRI study. Nat Neurosci 2(10):861–863
Gopnik M (1997) Language deficits and genetic factors. Trends Cogn Sci 1(1):5–9
Grice S, Spratling MW, Karmiloff-Smith A, Halit H, Csibra G, de Haan M, Johnson MH (2001) Disordered visual processing and oscillatory brain activity in autism and Williams Syndrome. Neuroreport 12:2697–2700
Grice SJ, de Haan M, Halit H, Johnson MH, Csibra G, Grant J, Karmiloff-Smith A (2003) ERP abnormalities of visual perception in Williams syndrome. Neuroreport 14:1773–1777
Hermer L, Spelke ES (1996) Modularity and development: the case of spatial reorientation. Cognition 61:195–232
Huttenlocher PR, Dabholkar AS (1997) Regional differences in synaptogenesis in human cerebral cortex. J Comp Neurol 387:167–178
Johnson MH (2001) Functional brain development in humans. Nat Rev Neurosci 2:475–483
Johnson MH, de Haan M (2001) Developing cortical specialization for visual-cognitive function: the case of face recognition. In: McClelland JL, Siegler RS (eds) Mechanisms of cognitive development: behavioral and neural perspectives. Lawrence Erlbaum Associates, UK, pp 253–270
Johnson MJ, Halit H, Grice SJ, Karmiloff-Smith A (2002) Neuroimaging and developmental disorders: a perspective from multiple levels of analysis. Dev Psychopathol 14:521–536
Karmiloff-Smith A (1992) Beyond modularity: a developmental approach to cognitive science. MIT Press, Cambridge, MA
Karmiloff-Smith A (1997) Crucial differences between developmental cognitive neuroscience and adult neuropsychology. Dev Neuropsychol 13(4):513–524
Karmiloff-Smith A (1998) Development itself is the key to understanding developmental disorders. Trends Cogn Sci 2(10):389–398
Karmiloff-Smith A (2006) The tortuous route from genes to behaviour: a neuroconstructivist approach. Cogn Affect Behav Neurosci 6(1):9–17
Karmiloff-Smith A (2007) Atypical Epigenesis. Dev Sci 10(1):84–88
Karmiloff-Smith A, Scerif G, Ansari D (2003) Double dissociations in developmental disorders? Theoretically misconceived, empirically dubious. Cortex 39:161–163

Karmiloff-Smith A, Thomas M, Annaz D, Humphreys K (2004) Exploring the williams syndrome face processing debate: the importance of building developmental trajectories. J Child Psychol Psychiatry 45(7):1258–1274

Kingsbury MA, Finlay BL (2001) The cortex in multidimensional space: where do cortical areas come from? Dev Sci 4(2):125–142

Leslie AM (1992) Pretense, autism, and the theory-of-mind module. Curr Dir Psychol Sci 1(1):18–21

Mareschal D, Johnson MH, Sirois S, Spratling M, Thomas M, Westermann G (2007) Neuroconstructivism. How the brain constructs cognition, vol 1. Oxford University Press, Oxford, UK

Mills DL, Alvarez TD, St. George M, Appelbaum LG, Bellugi U, Neville H (2000) Electrophysiological studies of face processing in williams syndrome. J Cogn Neurosci 12(suppl 1):47–64

Neville HJ, Mills DL, Bellugi U (1994) Effects of altered auditory sensitivity and age of language acquisition on the development of language – relevant neural systems: preliminary studies of Williams syndrome. In: Broman S, Grafman J (eds) Atypical cognitive deficits in developmental disorders: implications for brain function. Erlbaum, Hillsdale, NJ, pp 67–83

Piatelli-Palmerini M (2001) Speaking of learning. Nature 411:887–888

Pinker S (1994) The language instinct. Penguin Books, London

Spelke ES (1998) Nativism, empiricism, and the origins of knowledge. Infant Behav Dev 21(2):181–200

Spelke ES (2005) Physical knowledge in infancy: reflections on piaget's theory. In: Carey S, Gelman R (eds) The epigenesis of mind: essays on biology and cognition. Lawrence Erlbaum Associates, UK, pp 133–142

Sperber D (2001) In defense of massive modularity. In: Emmanuel D (ed) Language, brain, and cognitive development : essays in honor of jacques mehler. MIT Press, Cambridge, MA, pp 47–57

Temple CM (1997) Cognitive neuropsychology and its application to children. J Child Psychol Psychiatry 38(1):27–52

Udwin O, Yule W (1991) A cognitive and behavioural phenotype in Williams syndrome. J Clin Exp Neuropsychol 13:232–244

Young AW, Ellis HD (1989) Childhood Prosopagnosia. Brain Cogn 9(11):16–47

Theory of Mind

Beate Sodian and Susanne Kristen

Abstract The ability to attribute mental states to oneself and others is fundamental to human cognition and social behavior. Research on the development of a Theory of Mind in childhood indicates a two-step developmental sequence of desire-understanding and belief-understanding in preschool age. There is ongoing debate about the significance of recent findings on Theory of Mind in infancy. Neuroimaging studies of Theory of Mind reasoning in adults provide some support for a specific Theory of Mind network. This claim is contested, however, and many relevant studies have not yet been done. There is no hard evidence for a Theory of Mind (and an understanding of belief) in non-human primates, but there is evidence for a lower-level perception-goal psychology in some animals.

1 Introduction

A Theory of Mind is the ability to attribute mental states (thoughts, knowledge, beliefs, emotions, desires) to oneself and others. This common-sense mentalism is a powerful tool in our everyday predictions and explanations of human action. In developmental psychology, the child's conceptual understanding of the mental domain has been the focus of much research in the last 25 years (see Flavell 2004; Sodian 2005; Sodian and Thoermer 2006; Wellman 2002; for reviews). A critical test for the ability to represent mental states independently of reality is an understanding of *false* belief, since the ascription of true beliefs does not require a differentiation of beliefs from reality. Wimmer and Perner (1983) conducted the first systematic investigation of false belief understanding in children and found that children begin to correctly predict a story figure's mistaken action based on a false belief around the age of 4 years. In their classic "Maxi task" (1983) a doll named

B. Sodian (✉) and S. Kristen
Department of Psychology, Ludwig-Maximilians-Universität Munich,
Leopoldstr. 13, 80802, München, Germany
e-mail: sodian@edupsy.uni-muenchen.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_13, © Springer-Verlag Berlin Heidelberg 2010

Maxi is presented to the child and the experimenter tells the child that Maxi places his chocolate bar in the green cupboard. Maxi then goes to the playground. While he is away, a doll representing Maxi's mother wants to bake a cake. So she takes the chocolate out of the green cupboard, breaks off a piece and instead of putting it back into the green cupboard, puts it in the blue cupboard. The experimenter then describes how Maxi returns and his mother goes away again. After that, the child has to answer several questions. The memory questions relate to the child's understanding of the story. Basically all children are able to answer these correctly. However, the critical false belief question, relating to where Maxi will search for the chocolate, is the one that 3-year-olds generally do not pass. However, 40–80% (depending on the test condition) of 4- to 5-year-olds answer correctly that Maxi will search for the chocolate in the green cupboard. In contrast, younger children tend to make reality-based action predictions and fail to attribute false beliefs to other persons, as well as to themselves. Their general assumption is that Maxi will search for the chocolate in the blue cupboard. The nature and theoretical interpretation of this developmental phenomenon has attracted great interest in the past 25 years. More recently, Theory of Mind has also become a focus of neuroimaging research (see Amodio and Frith 2006; Saxe et al. 2004, for reviews). Furthermore, Theory of Mind development has been found to be related to the development of other cognitive functions such as language, memory, self-control, and time-representation (e.g., Astington and Jenkins 1999; Bischof-Köhler 2000; Perner et al. 2007; Perner and Lang 1999). Therefore, it is no longer possible to review all relevant lines of research in a brief chapter. In the following sections, we will briefly summarize developmental and neurocognitive Theory of Mind research, and then focus on the relation between Theory of Mind and language acquisition. Theory of Mind is not only an area of developmental psychology, but has also, from the start, focused on the question of whether non-human primates and other animals have mindreading abilities. We will conclude with a brief overview of recent progress in comparative Theory of Mind research.

2 Development of a Theory of Mind

Theory of Mind has also been described as a belief–desire psychology, since we rely on these two basic concepts in our everyday predictions and explanations of human action.

In child development, desire reasoning precedes belief reasoning by about one and a half years. Even18-month-old infants have a limited ability to reason non- egocentrically about people's desires, and by the age of two and a half years children make correct use of desire terms and grasp causal relations between desires and emotional outcomes. For example, they understand that people are happy when they get what they have desired (Bartsch and Wellman 1995).

In contrast, false belief understanding emerges only at the age of about 4 years. Three-year-olds and younger children fail to understand that a person's mental representation of reality can differ from reality, and they fail to understand how such

misrepresentations arise from false or incomplete information. Other conceptual distinctions that require an understanding of representational diversity are the appearance reality distinction, and the ability to understand that the same entity can be perceived differently from two different visual perspectives (Level 2 perspective taking). These distinctions are mastered in close conjunction with belief understanding around the age of 4 years (Flavell and Miller 1998). Consistent with the view that Theory of Mind development progresses as a two-step developmental sequence, Wellman and Liu (2004) found that tasks designed to assess children's understanding of desires, knowledge and beliefs form a Guttman scale. A meta-analysis of over 500 studies of false belief understanding showed that belief understanding is a robust developmental phenomenon. Although facilitating task conditions lead to success in children below the age of 4, there is still a clear developmental trend between the ages of about two and a half and 4 years (Wellman et al. 2001). Young children's difficulty with false belief tasks cannot be attributed to language demands, since non-verbal tasks have been shown to be equally difficult as verbal ones (Call and Tomasello 1999; Sodian et al. 2006); nor can it be attributed to inhibitory demands, since the developmental trend persists in tasks with low inhibitory demands, for instance tasks requiring an explanation for a mistaken action (Moses and Flavell 1990). There is evidence for a specific deficit in understanding *mental* representations in normally developing 3-year-olds and in autistic children (Leslie and Thaiss 1992; Perner et al. 1987).

An *implicit* understanding of belief precedes an explicit one by about 6 months (Clements and Perner 1994); also 36-month-olds take other people's false belief into account in communication (Carpenter et al. 2002). Recent eye-tracking studies have found evidence for belief-based anticipatory looking in infants as young as 24 months (Southgate et al. 2007), and 18 months (Neumann et al. 2008). Looking-time studies indicate that 13- and 15-month-old infants expect belief-based actions (Onishi and Baillargeon 2005; Surian et al. 2007). There is ongoing debate about whether these findings indicate that infants possess a Theory of Mind or whether they should be explained by lower-level heuristics, such as smart encoding or behavioral rules (e.g., Perner and Ruffman 2005). There is, however, undoubtedly, a rich understanding of goal-directed action in infancy, beginning around the age of 6 months (Woodward 1998). Around their first birthday, infants conceive of people as intentional agents (Tomasello 1999), paying attention to what other people are attending to and predicting their behavior from a variety of communicative cues. Infants use their intention-reading abilities in inferring others' goals even when the goal-directed action failed (Meltzoff 1995), in responding to bids for cooperation (Warneken and Tomasello 2006), and in distinguishing between unwillingness and inability of an adult to comply with their requests (Behne et al. 2005). Infants also encode what others see and do not see independently of their own visual access to an object (Luo and Baillargeon 2007; Sodian et al. 2007), and they use their knowledge of what others have seen in communication (Moll et al. 2007). Thus, recent research on infants' social understanding indicates that the preschooler's Theory of Mind is based on a rich understanding of intentional action in infancy. Longitudinal findings indicate that there is, in fact, a specific relation, on an individual level,

between infants' social information processing and preschoolers' Theory of Mind (Aschersleben et al. 2008; Thoermer et al. submitted; Wellman et al. 2004).

Later developments in children's understanding of the mind include second order false belief understanding around the age of 6 years, an increasingly powerful understanding of the mind as an active interpreter of information (Chandler and Carpendale 1998), which entails the notion of interpretive frameworks, rather than just simple beliefs. A nascent understanding of interpretive frameworks can be found even in 6-year-olds who understand the role of social prejudice in interpreting a target action (Pillow 1991). However, a full and explicit understanding of the role of theories or interpretive frameworks in interpreting natural phenomena develops slowly through adolescence and is not even present in all adults (Bullock et al. 2008). Later Theory of Mind development also includes elementary knowledge about thinking (Wellman et al. 1996). During the early preschool years, thinking is construed as an internal activity, representing a real or imagined content. However, an understanding of ongoing, constructive mental activity, and an intuitive idea of the stream of consciousness emerges only around the age of 8 years (Flavell 2003; Flavell and O'Donnell 1999). At this age, children understand that a person, seemingly unoccupied from the outside, e.g., just sitting on a bench, can still be preoccupied with mental activity on the inside.

3 Theories

There are several types of explanation for the development of children's knowledge about the mind. To date, the most dominant approach in philosophy and psychology is the so called Theory theory (Bartsch and Wellman 1995; Gopnik and Wellman 1994; Perner 1991; Wellman and Gelman 1998). Theory theorists offer an everyday, informal framework of related concepts as an explanation for mentalistic understanding. The developmental steps in these frameworks are analogous to the shift in scientific explanatory frameworks (Carey 1985). Bartsch and Wellman (1995) have described some of these critical steps, by arguing that 2-year-olds first develop a "desire psychology," basing their predictions about human behavior solely on desires. While desires remain the dominant explanations for people's behaviors at the next developmental stage, 3-year-olds begin to take beliefs into account and make use of a "desire-belief psychology." Finally, the relationship between beliefs and desires shifts and 4-year-old children understand that beliefs can be seen to "frame" desires and equally motivate human thought and behavior.

Perner (1991) developed an influential three stage model on children's developing representational skills. At the first stage, infants possess "primary representations," where they are limited to perceive things in current reality, e.g., a banana is a banana for them. During their second year of life, children entertain "secondary representations," which enable them to take "primary representations" and go beyond reality to model hypothetical situations, e.g., they pretend during child play that a banana represents a pistol. The third step is "metarepresentation." According to Perner (1991) older children's correct answers on false belief paradigms like the "Maxi task" (Wimmer and Perner 1983) are evidence of a true representational

understanding. To grasp the concept of false belief the child has to understand the difference between reality and a person's (false) concept about reality, but also that this false concept is believed to be true by the person. Thus, the child has to represent the representation of a representation. Theory theorists acknowledge that experiences play a major formative role in children's theory of mind.

Simulation theorists (Goldman 1992; Gordon 1986; Harris 1992) state that practice in role-play improves children's mentalising abilities, as children are enabled to understand other's mental states through role-taking and simulation processes.

Putting more emphasis on social experiences than cognitive theories of Theory of Mind, Carpendale and Lewis (2004) have introduced a social-constructivist approach. The basic idea is that, when socially interacting with other persons, children construe a Theory of Mind (Chapman 1991). Around the end of the first year of life, dyadic mother–infant face-to-face interactions are followed by triadic interactions between mother, child and object, allowing the child a gradual and cumulative acquisition of important mentalising abilities.

In contrast, other developmentalists (e.g., Carlson et al. 1998; Hughes 1998) believe that children's age-dependent improvement in a set of higher-order cognitive abilities, so-called executive functions, accounts for children's developing Theory of Mind skills.

As an example, until they are 4 years old most children fail the "windows task" devised by Russell et al. (1991). In this task the child is required to instigate a false belief in the experimenter. First, two boxes with transparent windows are presented, so that the child sees the chocolate reward in one of the boxes. Children are then required to infer the rule that when pointing at the empty box they can fool the experimenter and thus save the reward for themselves. They also have to realize that even though they know something to be false, someone else can be tricked into believing it to be true. Still, according to the executive function idea, younger children continually fail false belief tasks, as they lack the inhibitory control to suppress a prepotent response to the cognitively salient reality; in this case the reward in the box.

Modularity theorists (Baron-Cohen 1995; Leslie 1994; Scholl and Leslie 1999) postulate an acquisition of Theory of Mind through neurological processes. According to them, the maturation of a succession of domain-specific and modular mechanisms (Fodor 1983) enables organisms to deal with animate versus inanimate and agent versus nonagent objects. While the nature of these basic hard-wired mechanisms is not determined by experience, theorists in this field do not neglect the possibility that experience might trigger its operation and that its expression could be influenced by performance factors.

As the following section shows, neuroimaging studies have provided new insights into the existence of a brain region specialized in Theory of Mind.

4 Neural Correlates

A reliable set of brain regions has been connected with false belief reasoning, the marker test for Theory of Mind, including the medial prefrontal cortex (mPFC) (e.g., Goel et al. 1995; Sabbagh and Taylor 2000) and/or the right and left temporo-parietal

junction (TPJ) (Saxe and Kanwisher 2003). The few brain imaging studies with children have implicated activation of the mPFC (Kobayashi et al. 2007; Ohnishi et al. 2004), TPJ, inferior parietal lobule (Ohnishi et al. 2004) and ventral prefrontal cortex (Liu 2006). Evidence for a distinct Theory of Mind system would require a) increased brain activity for any task or stimulus eliciting the attribution of mental states and b) specialized processes, specifically devoted to Theory of Mind. This domain-specific interpretation of Theory of Mind would be challenged by the involvement of other processes like inhibitory control, language, executive function or recursion, serving a whole range of cognitive functions (domain-general processes). While developmental theorists (e.g., Wellman and Liu 2004) point out that Theory of Mind is to be understood as a complex ability consisting of more concepts than false belief, so far few neuroimaging studies have taken this into account. One study by Sommer et al. (2007) has compared true to false belief understanding. The results indicate that some Theory of Mind network regions, especially the right TPJ, are recruited only for false, not true belief attribution in adults. In line with Apperly et al. (2005) the results stress the importance of developing new tasks to isolate the distinct neural underpinnings of different mental concepts. Furthermore, fMRi studies, investigating the patterns of association and dissociation of deficits in patients with brain lesions and autistic children provide unique information concerning a distinct Theory of Mind network. Reviewing 20 years of data on lesion patients and children with autism, Stone and Gerrans (2006) argue that it may not be necessary to assume a separate Theory of Mind mechanism, since there is empirical evidence that Theory of Mind abilities do not solely depend on higher order cognitive processes or metarepresentation per se, but on their developmental and "online" interaction with low level precursor mechanisms like gaze processing and emotion recognition. This could explain why some studies show evidence that toddlers with autism have deficits in joint attention skills, but not always early deficits in executive function (Griffith et al. 1999; Rutherford and Rogers 2003). The fact that deficits in autistic children are not always apparent when they are tested by a computer rather than a person (Ozonoff 1995), hints at some indispensable input from lower order social domains to provide for intact higher order processes like executive function.

5 Theory of Mind and Language

Since it is not only Theory of Mind undergoing profound developmental changes during the first 5 years of life, but also children's language skills, and since Theory of Mind and language development have been found to be closely associated (Astington 2000), there is controversy about whether it is language ability that constrains Theory of Mind or vice versa (see Milligan et al. 2007, for a review). It has been shown that children's Theory of Mind assists them in their word learning (e.g., Baldwin 1991). In more complex communicative situations, adults' mentalizing ability was found to enhance the efficacy of shared understanding in conversation

(Krych-Appelbaum et al. 2007). For the reverse influence, it has been shown that parents' language about mental states facilitates children's later Theory of Mind and emotion understanding (e.g., Slaughter et al. 2007; Taumoepeau and Ruffman 2006). A longitudinal study by Astington and Jenkins (1999) provided evidence that early language ability predicts later Theory of Mind performance, while other studies (deVilliers and Pyers 2002; Slade and Ruffman 2005), have found the relation to be bi-directional. As a consequence, researchers' positions on the coevolution of language and Theory of Mind are fairly widespread. According to Ruffman (2000), since children's early Theory of Mind -components are of an implicit nature manifesting in children's overt behavior, rather than being insights they can consciously reflect and verbalize, it is statistical learning abilities (Saffran et al. 1996) that account for individual differences in early, nonverbal false belief understanding. Once this implicit understanding is in place, the first children to develop explicit understanding are those with better language skills because language provides the terms and means for refining implicit intuitions. Recent behavioral (Newton and deVilliers 2007) and neuroimaging studies (Kobayashi et al. 2007, 2008) of Theory of Mind development, indicating that adults process Theory of Mind more verbally than children, support this view. Interestingly, studies investigating the consequences of late acquired aphasia (especially loss of grammatical skills), suggest that a mature Theory of Mind functions even in the absence of syntactical structures and thus the neural bases of adult Theory of Mind and language might be largely distinct (Varley and Siegal 2000). As an example, an aphasia patient (Apperly et al. 2006), could still solve first and even second-order nonverbal Theory of Mind tasks. However, studies with autistic and normally developing children (Astington and Jenkins 1999; Lohmann and Tomasello 2003; Slade and Ruffman 2005; Tager-Flusberg and Sullivan 1994) indicate that comprehension of syntax is related to mentalising abilities. Accordingly, while syntax seems to be critical for developing a Theory of Mind, the structure and expression of mature, nonverbal belief reasoning might not depend on linguistic cues.

6 Theory of Mind in Other Species and Robots

Since Premack and Woodruff's (1978) seminal publication "Does the chimpanzee have a Theory of Mind ?" there has been a lively debate, especially in comparative psychology, whether and to what extent non-human animals can be credited with a Theory of Mind. Here again, a differentiated view on the different components of a Theory of Mind and its precursors seems crucial.

One of the building blocks for a Theory of Mind is the human infant's ability to follow gaze (see Emery 2000 for a review). While chimps are quite prolific gaze-followers their performance in respect to pointing, another important social cue, is mixed. While Call et al. (2000) and Barth et al. (2005) report positive responses to pointing and gazing, others found the responses to pointing to be very weak or not existing at all (i.e. Povinelli et al. 1997). Furthermore, other animals like ravens,

which are not as closely related to humans as chimps, follow a person's gaze into distant space (Bugnyar and Heinrich 2006). Though, in the *object choice-task* ravens, unlike chimpanzees (i.e. Call et al. 2000; Barth et al. 2005), do not rely on gaze cues to detect hidden food (Schloegl et al. 2007), indicating an ill-conceived understanding of the social function of gaze.

Another important precursor of Theory of Mind is intention understanding. With a paradigm Gergely et al. (2002) first tested on preverbal infants, Buttelmann et al. (2007b) found that like human infants, chimpanzees imitated an irrational action (switching a light on with one's head) more often, when it seemed necessary (the model's hands were blocked) compared to when it appeared as an act of free choice. Thus, to some extent, great apes seem to understand the intentionality and rationality of others' actions.

For the concept of "seeing" in chimps, Povinelli and Vonk (2003) suggest a behaviouristic rather than a mentalistic interpretation, while researchers from Tomasello's lab (Tomasello et al. 2003) advocate the idea that some mental states, "seeing" among them, can be understood to some extent by chimps. Karin D'Arcy and Povinelli (2002) found that, though chimpanzees in competitive feeding situations approach hidden food more often, this was independent of whether the food was behind a barrier blocking the rival's view or behind a barrier but in clear sight of a rival. Their results support the idea that chimps, while having competitive strategies, do not reason about what their conspecifics see or do not see.

An experiment by Bugnyar and Heinrich (2005) adds to the discussion by showing that ravens were able to know what other birds, competing about food with them, had or had not seen. While the authors conclude that ravens are candidates for the concept "see," they stress that they cannot rule out the possibility that the animals might have learned about another bird's viewpoint in relation to its later competitive behavior through foraging. Thus, they do not infer a full-fledged mentalistic understanding in ravens.

In a clever series of two studies, Buttelmann et al. (2007a) investigated whether chimpanzees use facial, emotional cues to infer the core concept of desire. In the first experiment, great apes were found to base their food- choice on the experimenter's emotional expression. In the second experiment, the chimpanzees first saw the experimenter lifting a cup and expressing a corresponding emotion of liking or disgust towards its content. Subsequently the animals' view was blocked. Without having visual access as to which cup exactly the experimenter has lifted, the animals saw the happy-looking experimenter eating food out of one of the containers. After that the chimpanzees could choose one of the cups for themselves. Chimpanzees more often chose the cup the experimenter had expressed disgust towards, obviously inferring that this would be the one still containing food. Thus, chimpanzees seem to understand other's desires and based on that, can make some action predictions; in this case that the experimenter had eaten the food he desired. It is yet to be investigated whether chimpanzees understand the subjective quality of desires and, like 18-month-old human infants, differentiate between their own and another person's desire (Repacholi and Gopnik 1997).

Call and Tomasello (2008) recently reviewed 30 years of research and concluded that while chimpanzees can infer the goals and intentions of others and grasp the concepts of perception and knowledge, there is no evidence that they possess any false belief understanding comparable to humans. As a consequence, while Bartsch and Wellman (1995) have proposed a belief-desire-theory for human's Theory of Mind, they propose a perception-goal philosophy for the primate's understanding of the mental world.

The field of Theory of Mind will further emerge and seek input from other disciplines. While philosophers, neuroscientists and ethologists have jointly contributed to Theory of Mind research, robotics is a newly emerging area adding to the field. To build a humanoid robot that can participate in social interaction, scientists in robotics have to address the same issues as researchers of social cognition. The benefit could be bi-directional though. Scassellati (2002), who has performed research in this area at the MIT Artificial intelligence lab, points out several advantages of applying robotics as a tool for cognitive science. As an example, the validity and predictive powers of theoretical models of a Theory of Mind could be tested against each other by manipulating the robot in a controlled and detailed way, while maintaining the same setting and testing paradigms as with human subjects. By varying internal model parameters, one could systematically study environmental effects on each step of Theory of Mind development. Furthermore, a humanoid robot could be subjected to controversial testing, which would be unethical, expensive or too dangerous to perform on human subjects.

As a first step, Scassellati (2002) has discussed the module theories of Leslie (1994) and Baron-Cohen (1995) in the realms of robotics. More concretely, he has developed initial implementation details of basic mind reading skills in robots (e.g., tracking human faces and eyes and differentiating inanimate from animate objects). What thus unites researchers of infant social cognition and researchers constructing humanoid robots is that both fields are based on a careful conceptual analysis and profound theory building as prerequisites for critical empirical examinations.

References

Amodio DM, Frith CD (2006) Meetings of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268–277

Apperly IA, Samson D, Humphreys GW (2005) Domain-specifity and theory of mind: evaluating neuropsychological evidence. Trends Cogn Sci 9:572–577

Apperly IA, Samson D, Carroll N, Hussain S, Humphreys G (2006) Intact first-and second-order false belief reasoning in a patient with severely impaired grammar. Soc Neurosci 1:334–348

Aschersleben G, Hofer T, Jovanovic B (2008) The link between infant attention to goal-directed action and later theory of mind abilities. Dev Sci 11:862–868

Astington JW (2000) Language and metalanguage in children's understanding of mind. In: Zelazo PD, Astington JW (eds) Minds in the making: essay in honor of David R Olson Malden. Blackwell, MA, pp 267–284

Astington JW, Jenkins JM (1999) A longitudinal study on the relation between language and theory of mind development. Dev Psychol 35:1311–1320

Baldwin D (1991) Infants' contribution to the achievement of joint reference. Child Dev 62:875–890
Baron-Cohen S (1995) Mindblindness: an essay on autism and theory of mind. MIT Press, Cambridge, MA
Barth J, Reaux JE, Povinelli DJ (2005) Chimpanzees' (Pan troglodytes) use of gaze cues in object-choice tasks: different methods yield different results. Anim Cogn 8:84–92
Bartsch K, Wellman HM (1995) Children talk about the mind. University Press, Oxford
Behne T, Carpenter M, Call J, Tomasello M (2005) Unwilling versus unable: infants' understanding of intentional action. Deve Psychol 41:328–337
Bischof-Köhler D (2000) Kinder auf Zeitreise. Theory of Mind, Zeitverständnis und Handlungsorganisation (Children's mental time travel. Theory of mind, concept of time and organisation of behavior). Bern: Hans Huber
Bugnyar T, Heinrich B (2005) Food-storing ravens differentiate between knowledgable and ignorant competitors. Proc R Soc Lond B 272:1641–1646
Bugnyar T, Heinrich B (2006) Pilfering ravens, Corvus corax, adjust their behavior to social context and identity of competitors. Anim Cogn 9:369–376
Bullock M, Sodian B, Koerber S (2008) Doing experiments and understanding science. Development of scientific reasoning from childhood to adulthood. In: Schneider W, Bullock M (eds) Human development from early childhood to early adulthood: findings from a 20 Year longitudinal study. Erlbaum, Mahwah, NJ
Buttelmann D, Call J, Tomasello M (2007) Great apes' referential use of emotional expressions. In: Poster presented at the 37th Annual Meeting of the Jean Piaget Society, "Developmental Social Cognitive Neuroscience", Amsterdam, The Netherlands
Buttelmann D, Carpenter M, Call J, Tomasello M (2007b) Enculturated chimpanzees imitate rationally. Dev Sci 10:F31–F38
Call J, Tomasello M (1999) A nonverbal false belief task: the performance of children and great apes. Child Dev 70:381–395
Call J, Tomasello M (2008) Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci 12:187–192
Call J, Agnetta B, Tomasello M (2000) Cues that chimpanzees do and do not find hidden objects. Anim Cogn 3:23–34
Carey S (1985) Conceptual change in childhood. Cambridge University Press, Cambridge, MA
Carlson SM, Moses LJ, Hix H (1998) The role of inhibitory processes in young children's difficulties with deception and false belief. Child Dev 69:672–691
Carpendale JJ, Lewis C (2004) Constructing an understanding of mind: the development of children's social understanding within social interaction. Behav Brain Sci 27:79–151
Carpenter M, Call J, Tomasello M (2002) A new false belief test for 36-month-olds. Br J Dev Psychol 20:393–420
Chandler MJ, Carpendale JIM (1998) Inching toward a mature theory of mind. In: Ferrari M, Sternberg RJ (eds) Self-awareness: its nature and development. Guilford Press, New York, pp 148–190
Chapman M (1991) The epistemic triangle: operative and communicative components of cognitive competence. In: Chandler M, Chapman M (eds) Criteria for competence: controversies in the conceptualization and assessment of children's abilities. Erlbaum, Hillsdale, NJ, pp 209–228
Clements WA, Perner J (1994) Implicit understanding of belief. Cogn Dev 9:377–395
deVilliers JG, Pyers JE (2002) Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief-understanding. Cogn Dev 17:1037–1060
Emery NJ (2000) The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev 24:581–604
Flavell JH (2003) 2003 Heinz Werner Lecture Series: Vol. 25. Development of children's knowledge about the mind. Clark University Press, Worcester, MA
Flavell JH (2004) Theory-of-mind development: retrospect and prospect. Merrill-Palmer Q 50:274–290

Flavell JH, Miller PH (1998) Social cognition. In: Kuhn D, Siegler RS (eds) Handbook of child psychology. Volume 2: cognition, perception and language, vol 5. Wiley, New York, pp 851–898
Flavell JH, O'Donnell AK (1999) Le développement de savoirs intuitifs à propos des expériences mentales (Development of intuitions about mental experiences). Enfance 51:267–276
Fodor JA (1983) Modularity of mind. MIT Press, Cambridge, MA
Gergely G, Bekkering H, Király I (2002) Rational imitation in preverbal infants. Nature 415:755
Goel V, Grafman J, Sadato N, Hallett M (1995) Modeling other minds. Neuroreport 6:1741–1746
Goldman AI (1992) In defense of the simulation theory. Mind Lang 7:104–119
Gopnik A, Wellman HM (1994) The theory theory. In: Hirschfeld LA, Gelman SA (eds) Mapping the mind-domain specificity in cognition and culture. Cambridge University Press, Cambridge, pp 257–293
Gordon RM (1986) Folk psychology as simulation. Mind Lang 1:158–171
Griffith EM, Pennington BF, Wehner EA, Rogers S (1999) Executive functions in young children with autism. Child Dev 70:817–832
Harris PL (1992) From simulation to folk psychology: the case for development. Mind Lang 7:120–144
Hughes C (1998) Executive function in preschoolers: links with theory of mind and verbal ability. Br J Dev Psychol 16:233–253
Karin D'Arcy MR, Povinelli DJ (2002) Do chimpanzees know what each other see? A closer look. Int J Comp Psychol 15:21–54
Kobayashi C, Glover GH, Temple E (2007) Children's and adults' neural bases of verbal and nonverbal 'Theory of Mind'. Neuropsychologia 45:1522–1532
Kobayashi C, Glover GH, Temple E (2008) Switching languages switches mind: linguistic effect on developmental neural bases of 'Theory of mind'. SCAN 8:62–70
Krych-Appelbaum M, Law JB, Jones D, Barnacz A, Johnson A, Keenan JP (2007) I think I know what you mean: the role of theory of mind in collaborative communication. Interact Stud 8:267–280
Leslie AM (1994) ToMM, ToBY, and agency: core architecture and domain specifity. In: Hirschfeld LA, Gelman SA (eds) Mapping the mind: domain specifity in cognition and culture. Cambridge University Press, Cambridge, pp 119–148
Leslie AM, Thaiss L (1992) Domain-specificity in conceptual development: evidence from autism. Cognition 43:225–251
Liu D (2006) Neural correlates of children's theory of mind development. In: Psychology, Ph. D. Thesis. Ann Arbor University of Michigan, MI
Lohmann H, Tomasello M (2003) The role of language in the development of false belief understanding: a training study. Child Dev 74:1130–1144
Luo Y, Baillargeon R (2007) Do 12.5 -month-old infants consider what objects others can see when interpreting their actions? Cognition 105:489–512
Meltzoff AN (1995) Understanding the intentions of others: re-enactment of intended acts by 18-month-old-children. Dev Psychol 31:838–850
Milligan K, Astington JW, Dack LA (2007) Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. Child Dev 78:622–646
Moll H, Carpenter M, Tomasello M (2007) Fourteen-month-old infants know what others experience in jointly engagement with them. Dev Sci 10:826–835
Moses LJ, Flavell JH (1990) Inferring false beliefs from actions and reactions. Child Dev 61:929–945
Neumann A, Thoermer C, Sodian B (2008) Can 18-month-olds overcome the reality bias in theory of mind tasks? New evidence from eye-tracking. In: Poster presented at the International Conference of Infant Studies, Vancouver, Canada
Newton AM, deVilliers JG (2007) Thinking while talking. Psychol Sci 18:574–579
Ohnishi T, Moriguchi Y, Matsuda H, Mori T, Hirakata M, Imabayashi E et al (2004) The neural network for the mirror system and the 'theory of mind' in normally developed children: an fMRI Study. NeuroReport 15:1483

Onishi K, Baillargeon R (2005) Do 15-month-old infants understand false beliefs? Science 308:255–258
Ozonoff S (1995) Reliability and validity of the wisconsin card sorting test in studies of autism. Neuropsychology 9:491–500
Perner J (1991) Understanding the representational mind. MIT Press, Cambridge, MA
Perner J, Lang B (1999) Development of theory of mind and executive control. Trends Cogn Sci 3:337–344
Perner J, Ruffman T (2005) Infants' insight into the mind: how deep. Science 308:214–216
Perner J, Kloo D, Gornik E (2007) Episodic memory development: theory of mind is part of re-experiencing experienced events. Infant Child Dev 16:471–490
Perner J, Leekam SR, Wimmer H (1987) Three-year old's difficulty with false belief: the case for a conceptual deficit. Br J Dev Psychol 5:125–137
Pillow BH (1991) Children's understanding of biased social cognition. Dev Psychol 27:539–551
Povinelli DJ, Vonk J (2003) Chimpanzee minds: suspiciously human? Trends Cogn Sci 7:157–160
Povinelli DJ, Reaux JE, Bierschwale DT, Allain AD, Simon BB (1997) Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. Cogn Dev 12:423–461
Premack DG, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1:515–526
Repacholi B, Gopnik A (1997) Early understanding of desires: evidence from 14 and 18-month-olds. Dev Psychol 33:12–21
Ruffman T (2000) Nonverbal theory of mind: is it important, is it implicit, is it simulation, is it relevant to autism? In: Astington JW (ed) Minds in the making: essays in honor of David R.Olson. Blackwell, Oxford, pp 250–266
Russell J, Mauthner N, Sharpe S, Tidswell T (1991) The "windows task" as a measure of strategic deception in preschoolers and autistic subjects. Br J Dev Psychol 9:331–349
Rutherford M, Rogers SJ (2003) Cognitive underpinnings of pretend play in autism. J Autism Dev Disord 33:289–302
Sabbagh MA, Taylor M (2000) Neural correlates of theory-of-mind reasoning: an event related potential study. Psychol Sci 11:46–50
Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. Science 274:1926–1928
Saxe R, Kanwisher N (2003) People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". Neuroimage 19:1835–1842
Saxe R, Carey S, Kanwisher N (2004) Understanding other minds: linking developmental psychology and functional neuroimaging. Annu Rev Psychol 55:87–124
Scassellati B (2002) Theory of mind for a humanoid robot. Auton Robot 12:13–24
Schloegl C, Kotschral K, Bugnyar T (2007) Gaze following in common ravens, Corvus corax: ontongeny and habituation. Anim Behav 74:769–778
Scholl BJ, Leslie AM (1999) Modularity, development, and "theory of mind". Mind Lang 14:131–153
Slade L, Ruffman T (2005) How language does (and does not) relate to theory of mind: a longitudinal study of syntax, semantics, working memory and false belief. Br J Dev Psychol 23:1–26
Slaughter V, Peterson CC, Mackintosh E (2007) Mind what mother says: narrative input and theory of mind in typical children and those on the autism spectrum. Child Dev 78:839–858
Sodian B (2005) Theory of mind. The case for conceptual development. In: Schneider W, Schumann-Hengsteler R, Sodian B (eds) Interrelations among working memory, theory of mind, and executive functions. Erlbaum, Hillsdale, NJ, pp 95–130
Sodian B, Thoermer C (2006) Theory of mind. In: Schneider W, Sodian B (eds) Enzyklopädie der Psychologie, Themenbereich C, Serie V, Band 2: Kognitive Entwicklung. Hogrefe, Göttingen, pp 495–608
Sodian B, Thoermer C, Dietrich N (2006) Two-to-four-year-old children's differentiation of knowing and guessing in a non-verbal task. Eur J Dev Psychol 3:222–237
Sodian B, Thoermer C, Metz U (2007) Now I see but you don't: 14-month-olds can represent another person's visual perspective. Dev Sci 10:199–204

Sommer M, Doehnel K, Sodian B, Meinhardt J, Thoermer C, Hajak G (2007) Neural correlates of true and false belief reasoning. Neuroimage 35:1378–1384
Southgate V, Senju A, Csibra G (2007) Action anticipation through attribution of false belief by two-year-olds. Psychol Sci 18:587–592
Stone VE, Gerrans P (2006) What's domain-specific about theory of mind? Soc Neurosci 1:309–319
Surian L, Caldi S, Sperber D (2007) Attribution of beliefs by 13-month-old infants. Psychol Sci 18:580–586
Tager-Flusberg H, Sullivan K (1994) A second look at second-order belief attribution in autism. J Autism Dev Disord 24:577–586
Taumoepeau M, Ruffman T (2006) Mother and infant mental state talk relates to desire language and emotion understanding. Child Dev 77:465–481
Thoermer C, Sodian, B, Nickelt J (submitted). Understanding the pointing gesture at 14 months predicts Theory of Mind at four years. Unpublished Ms. University of Munich
Tomasello M (1999) Having intentions, understanding intentions and understanding communicative intentions. In: Zelazo PD, Astington JW, Olson DR (eds) Developing theories of intention. Erlbaum, Mahwah, NJ, pp 63–75
Tomasello M, Call J, Hare B (2003) Chimpanzees vs. humans: it's not that simple. Trends Cogn Sci 7:239–240
Varley R, Siegal M (2000) Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. Curr Biol 10:723–726
Warneken F, Tomasello M (2006) Altruistic helping in human infants and young chimpanzees. Science 3:1301–1303
Wellman H, Liu D (2004) Scaling of theory-of-mind tasks. Child Dev 75:523–541
Wellman HM (2002) Understanding the psychological world: developing a theory of mind. In: Goswami U (ed) The Blackwell handbook of childhood cognitive development. Blackwell, Oxford, pp 167–187
Wellman HM, Gelman SA (1998) Knowledge acquisition in foundational domains. In: Kuhn D, Siegle RS (eds) Handbook of child psychology, Vol 2: cognition, perception and language, 5th edn. Wiley, New York, pp 523–573
Wellman HM, Cross D, Watson J (2001) Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev 72:655–684
Wellman HM, Hollander M, Schult CA (1996) Young children's understanding of thought-bubbles and of thoughts. Child Dev 67:768–788
Wellman HM, Phillips AT, Dunphy-Lelii S, LaLonde N (2004) Infant social attention predicts preschool social cognition. Dev Sci 7:283–288
Wimmer H, Perner J (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13:103–128
Woodward AL (1998) Infants selectively encode the goal object of an actor's reach. Cognition 69:1–34

The Development of Metacognitive Competencies

Wolfgang Schneider

Abstract This paper describes historical and current trends in research on the development of metacognitive competencies. Stimulated by classic theoretical analyses of the concept of metacognition initiated by Ann Brown, John Flavell and their colleagues, contemporary extensions of the concept emphasize the important roles of both procedural and declarative metacognition for successful information processes. Major research findings on the development of these two components of metacognition are reviewed, and links between children's early "theory of mind" and subsequent verbalizable metamemory are described. Next, new evidence on children's metacognitive development in childhood and adolescence is summarized, indicating major shifts in children's declarative metacognitive knowledge, in particular, their strategy knowledge, between the end of kindergarten and the end of elementary school. Although similarly fast developments could not be demonstrated for procedural metacognitive knowledge, several empirical studies suggest developmental changes in the relationship between monitoring and self-regulatory abilities, with older (but not younger) children being able to regulate their achievement-related behavior based on the outcome of their monitoring attempts. Finally, the paper reviews classic and contemporary applications of metacognitive theory to various educational settings, generally illustrating the importance of metacognition for various aspects of academic performance.

1 Introduction

Research on the development of "metacognition" was initiated in the early 1970s by Ann Brown, John Flavell and their colleagues (for reviews, see Brown et al. 1983; Flavell et al. 2002; Goswami 2008; Schneider and Pressley 1997). Although various definitions of the term "metacognition" have been used in the literature on

W. Schneider
Department of Psychology, University of Würzburg, Wittelsbacherplatz 1, 97074, Würzburg, Germany
e-mail: schneider@psychologie.uni-wuerzburg.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_14,

cognitive development, the concept has usually been broadly and rather loosely defined as any knowledge or cognitive activity that takes as its object, or regulates, any aspect of any cognitive enterprise (cf. Flavell et al. 2002). Obviously, this conceptualization refers to people's knowledge of their own information processing skills, as well as knowledge about the nature of cognitive tasks, and about strategies for coping with such tasks. Moreover, it also includes executive skills related to monitoring and self-regulation of one's own cognitive activities. In a seminal paper, Flavell (1979) described three major facets of metacognition, namely metacognitive knowledge, metacognitive experiences, and metacognitive skills, that is, strategies controlling cognition. According to Flavell et al. (2002), declarative metacognitive knowledge refers to the segment of world knowledge that concerns the human mind and its doings. For instance, metacognitive knowledge about memory includes explicit, conscious, and factual knowledge about the importance of person, task, and strategy variables for memorizing and recalling information. A person is said to possess "conditional" metacognitive knowledge whenever he or she is able to justify or explain the impact of person, task, and strategy variables on memory performance (see Paris and Oka 1986). Metacognitive experiences refer to a person's awareness and feelings elicited in a problem-solving situation (e.g., feelings of knowing), and metacognitive skills are believed to play a role in many types of cognitive activity such as oral communication of information, reading comprehension, attention, and memory. These facets of metacognition refer to a person's procedural knowledge, which Brown and colleagues (1983) referred to as "knowing how", and which can be further subdivided in monitoring and self-regulatory functions (see below). For an excellent discussion of more subtle distinctions among various aspects of metacognition, see Kuhn (1999, 2000).

This theoretical framework of metacognition was subsequently extended by Pressley, Borkowski, and their colleagues (e.g., Pressley et al. 1989), who proposed an elaborate model of metacognition, the Good Information Processing Model, that not only considered aspects of procedural and declarative metacognitive knowledge but also linked these concepts to other features of successful information processing. According to this model, sophisticated metacognition is closely related to the learner's strategy use, domain knowledge, motivational orientation, general knowledge about the world, and automated use of efficient learning procedures. All of these components are assumed to interact. For instance, specific strategy knowledge influences the adequate application of metacognitive strategies, which in turn affects knowledge. As the strategies are carried out, they are monitored and evaluated, which leads to expansion and refinement of specific strategy knowledge.

It should be noted that conceptualizations of metacognition developed in the fields of general psychology, social psychology, and the psychology of aging typically differ from this taxonomy. Popular conceptualizations of metacognition in the field of cognitive psychology exclusively elaborate on the procedural component, focusing on the interplay between monitoring and self-control (see Nelson 1996). On the other hand, when issues of declarative metacognitive knowledge are analyzed in the fields of social psychology and gerontology, the focus is on a person's belief about cognitive phenomena and not on veridical knowledge.

More recent conceptualizations of metacognition added components such as self-regulation skills (e.g., Efklides 2001; Schunk and Zimmerman 1998). While the concept of metacognition was first developed in the context of developmental research, it is now widely used in different areas of psychology, including motivation research, clinical and educational psychology. Recent developments also include cognitive neuroscience models of metacognition (cf. Shimamura 2000). Its popularity is mainly due to the fact that metacognition is crucial for concepts of everyday reasoning and those assessing scientific thinking as well as social interactions.

2 Classic Research on the Development of Metamemory

2.1 Declarative Metamemory

From the very beginning, research on the development of metacognitive knowledge has focused on the domain of memory. Flavell and Wellman (1977) coined the term "metamemory" to refer to children's knowledge about what memory is, how it works, and which factors influence its functioning. Using sensitive methods that minimize demands on the child, it has been possible to demonstrate some rudimentary knowledge about memory functioning in preschoolers. Knowledge of facts about memory develops impressively during the course of elementary school and beyond, reaching its peak in late adolescence and young adulthood (cf. Schneider and Lockl 2002). It seems important to note that even though metacognitive knowledge increases substantially between young childhood and young adulthood, there is also evidence that many adolescents (including college students) demonstrate little knowledge of powerful and important memory strategies when the task is to read, comprehend, and memorize complex text materials (cf. Brown et al. 1983; Garner 1987; Pressley and Afflerbach 1995). Also, knowledge about possible interactions among memory variables (e.g., task demands and strategies) seems to develop rather late and continues to improve after adolescence (Schneider and Pressley 1997).

Taken together, the empirical evidence demonstrates that some declarative knowledge is already available in preschoolers and kindergarten children, and that this component of metamemory develops steadily over the elementary school years and beyond. Nonetheless, declarative metamemory is not complete by the end of childhood.

2.2 Procedural Metamemory

Several early metamemory studies explored how children use their metacognitive knowledge to monitor and self-regulate their mnemonic activities. While self-monitoring involves knowing where you are with regard to your goal of understanding and memorizing task materials, self-regulation includes planning, directing, and evaluating

one's mnemonic activities (cf. Flavell et al. 2002). Early research focusing on monitoring showed that even young children seem to possess the relevant skills, particularly when the memory tasks were not very difficult (see the review by Schneider and Pressley 1997). However, the evidence regarding developmental trends was not consistent, with some studies showing better performance in younger than in older children, and others illustrating age-correlated improvement.

2.3 Metamemory–Memory Relations

From a developmental and educational perspective, the metamemory concept seems well-suited to explain children's "production deficiencies" on a broad variety of memory tasks. Early empirical research on metamemory was stimulated by the belief that young children do not spontaneously use memory strategies because they are not familiar with memory tasks and are unable to judge the advantages of memory strategies such as rehearsal or categorization. Metamemory researchers assumed that this situation should change after children enter school and are confronted with numerous memory tasks. Experience with such tasks should improve strategy knowledge, which in turn should exert a positive influence on memory behavior (e.g., strategy use). Thus, a major motivation behind studying metamemory and its development was the assumption that although links between metamemory and memory may be weak in early childhood, they should become much stronger with increasing age.

Overall, the empirical findings do not indicate a very strong relationship, even though the numbers show reliable associations. For instance, a statistical meta-analysis of 60 studies (with more than 7,000 participants) produced an average correlation of 0.41 (Schneider and Pressley 1997, p. 220). The size of the correlation seems to depend on factors such as type of task, age of children, task difficulty, and timing of metamemory assessment (before or after the memory task). The causal relation between metamemory and memory is also complex in that metamemory sometimes has an indirect effect on recall, as when knowledge about categorization strategies leads to semantic grouping during the study period, which in turn produces better recall. Moreover, the influence seems to be bidirectional (cf. Flavell et al. 2002; Hasselhorn 1990). That is, metamemory can influence memory behavior, which in turn leads to enhanced metamemory.

3 Development of Metacognitive Knowledge and "Theory of Mind"

Given that this chapter focuses on the development of metacognitive knowledge, it seems important to elaborate on the differences between the classic older metamemory research paradigm and more recent theory-of-mind research (see also Flavell 2000; Kuhn 1999, 2000). While metacognitive development has been

studied more in terms of the important mechanisms operating within individual minds, exploring children's awareness of their own cognition, theory-of-mind (ToM) research is concerned with what children know about somebody else's mind (cf. Goswami 2008; Kuhn 1999; Schneider and Lockl 2002). Another distinction between the two research paradigms concerns the age groups under study. Because ToM researchers are mainly interested in the origins of knowledge about mental states, they predominantly study infants and young children. On the other hand, metacognitive researchers investigate knowledge components and skills that already require some understanding of mental states, and thus mainly test older children and adolescents. Despite this difference in focus, these two research paradigms are connected in important ways. One of the first to detect this relationship was Wellman (1985), who suggested that metacognition consists of a "large, multi-faceted theory of mind" (p. 29).

An influential recent research paradigm has aimed at understanding metacognitive processes in their developmental dimension, trying to link young children's "theory of mind" with their subsequent metacognitive developments. The most important aspects of this work will be summarized next.

3.1 Assessment of Children's "Theory-of-Mind"

From the early 1980s on, numerous studies have focused on young children's knowledge about the mental world, dealing with very young children's understanding of mental life and age-related changes in this understanding, for instance, their knowledge that mental representations of events need not correspond to reality (cf. Perner 1991; Wellman 1985). One of the major and consistent outcomes of these ToM studies has been that significant changes in children's ability to take over the perspective of other people occur between 3 and 4 years of age. Explanations of this rapid change in children's ToM were linked to developmental changes in functions of the prefrontal cortex, in particular, inhibitory functions and those concerned with the regulation of behavior.

3.2 Links Between Theory of Mind and Metacognitive Knowledge

Several years ago, a longitudinal study was started in our lab with 174 children (who were about 3 years of age at the beginning) that investigated the relationship between early theory of mind and subsequent metamemory development, while simultaneously taking into account the possible mediating role of language development (for more details, see Lockl and Schneider 2006, 2007). Children were tested at four measurement points, separated by a testing interval of approximately half a year. While the main goal was to combine aspects of research on ToM and metamemory within a longitudinal framework, a second goal was to

examine the role of language abilities in the emergence of theory-of-mind and metacognitive competencies.

There were several interesting findings. First of all, we demonstrated rapid improvements in both language competencies and children's theory of mind over the age period under study, this confirming previous longitudinal research on this issue (see Astington and Jenkins 1999; Schneider et al. 1999). Secondly, we were able to show that the stability of the theory of mind construct was only moderate at the beginning but increased subsequently, reaching levels of stability similar to those found for the language tests. This finding clearly points to a continuity in ToM development.

Furthermore, several outcomes addressing the impact of language on ToM and metamemory development seem notable. Findings demonstrated a strong relationship between language and theory of mind, thus confirming results of previous studies (e.g., Ruffman et al. 2002). Moreover, significant relationships between language and metamemory could be shown. That is, language abilities assessed at the ages of 3 and 4 years made significant contributions to the prediction of metamemory scores at the age of 5. Finally, it was shown that theory of mind obviously facilitated the acquisition of metacognitive knowledge. While the amount of variance in metamemory scores at the age of 5 explained by ToM at the age of 3 was relatively small, this proportion increased considerably when ToM scores assessed at age 4 were used as predictors. Early ToM competencies also affected the acquisition of metacognitive vocabulary (e.g., knowledge about mental words such as guessing or knowing), which in turn had an impact on developmental changes in metacognitive knowledge. Obviously, advanced ToM development is characterized by a growing insight into inferential and interpretive mental processes (Sodian 2005). Overall, we demonstrated that children who acquired a theory of mind early also showed better metamemory performance assessed about 2 years later. These findings support the hypothesis that early ToM competencies can be considered as a precursor of subsequent metamemory.

4 New Evidence Concerning Metacognitive Development in Childhood and Adolescence

As already noted above, children's declarative metamemory increases with age and is correlated with age-related improvements in memory behavior (see Schneider and Lockl 2002; Schneider and Pressley 1997, for reviews). We know from various interview studies that knowledge about memory-relevant knowledge concerning person, task, and strategy variables develops significantly from the early elementary school period on and does not reach its peak before young adulthood (cf. Schneider and Pressley 1997). For instance, factual knowledge about the importance of task characteristics and memory strategies develops rapidly once children enter school. Knowledge about the usefulness of memory strategies was tapped in several studies that focused on organizational strategies (e.g., Justice 1985; Schneider 1986; Sodian et al. 1986). As a main result, these studies reported a major shift in strategy

knowledge between kindergarten and Grade 6, a finding replicated in numerous recent studies (e.g., Schneider et al. in press).

Taken together, recent studies on declarative metacognitive knowledge more or less confirmed outcomes of previous research. In comparison, more recent investigations on procedural metacognitive knowledge and its development produced several new insights concerning developmental trends and will be discussed in some detail below.

4.1 The Development of Self-Monitoring and Self-Control

According to Nelson and Narens (1990, 1994), self-monitoring and self-regulation correspond to two different levels of metacognitive processing that interact very closely. Self-monitoring refers to keeping track of where you are with your goal of understanding and remembering (a bottom-up process). In comparison, self-regulation or control refers to central executive activities and includes planning, directing, and evaluating your behavior (a top-down process).

4.2 Monitoring Skills in Children

The most studied type of procedural metamemory is that of self-monitoring, evaluating how well one is progressing (cf. Borkowski et al. 1988; Brown et al. 1983; Schneider and Lockl 2002). The developmental literature has focused on monitoring components such as ease-of-learning (EOL) judgments, judgments of learning (JOL), and feeling-of-knowing (FOK) judgments. What are the major developmental trends? In short, the findings suggest that even young children possess monitoring skills, and that developmental trends are not entirely clear, varying as a function of the paradigm under study. While young kindergarten children tend to overestimate their performance when EOL judgments are considered, EOL judgments can be already accurate in young elementary school children. In most of the relevant studies, subtle improvements over the elementary school years were found (cf. Schneider and Lockl 2002, in press).

Given that only a few developmental studies focused on judgments of learning (JOLs) occurring during or soon after the acquisition of memory materials, the situation is not yet clear. Overall, findings support the assumption that children's ability to judge their own memory performance after study of test materials seems to increase over the elementary school years. However, even young children are able to monitor their performance quite accurately when judgments are not given immediately after study but are somewhat delayed.

A number of studies have explored children's feeling-of-knowing (FOK) judgments and accuracy (e.g., Cultice et al. 1983; DeLoache and Brown 1984). FOK judgments occur either during or after a learning procedure and are judgments about whether

a currently unrecallable item will be remembered at a subsequent retention test. Typically, children are shown a series of items and asked to name them. When children cannot recall the name of an object given its picture, they are asked to indicate whether the name could be recognized if the experimenter provided it. These FOK ratings are then related to subsequent performance on the recognition test. Overall, most of the available evidence on FOK judgments suggests that FOK accuracy improves continuously across childhood and adolescence (e.g., Wellman 1977; Zabrucky and Ratner 1986).

A more recent study by Lockl and Schneider (2002) using the same experimental paradigm was in accord with the classic findings described above. One of the aims of this study was to explore the basis of FOK judgments by comparing the traditional "trace-based" view with the "trace accessibility mode" developed by Koriat (1993). While the former assumes a two-stage process of monitoring and retrieval, the latter proposes that FOK judgments are based on retrieval attempts and determined by the amount of information that can be spontaneously generated, regardless of its correctness. One important prediction derived from this model is that FOK judgments for correctly recalled and incorrect answers (commission errors) should be comparably high, and also higher than FOK judgments for omission errors. This is what Lockl and Schneider (2002) actually found: While the magnitude of FOK judgments given after Commission errors did not differ much from those provided after correct recall and was significantly higher than that given after omission errors, recognition performance was comparable in the case of commission and omission errors. Thus the assumption that feeling of knowing can be dissociated from knowing was empirically confirmed.

Taken together, recent research assessing monitoring abilities in JOL or FOK tasks demonstrates rather small developmental progression in children's monitoring skills (see also Roebers et al. 2007).

4.3 The Relation Between Monitoring and Control Processes in Children

An important reason to study metacognitive monitoring processes is because monitoring is supposed to play a central role in directing how people study. Numerous studies including adult participants showed that individuals use memory monitoring, especially judgments of learning (JOLs), to decide which items to study and how long to spend on them (e.g., Metcalfe 2002; Nelson and Narens 1990). However, little is known about how children use monitoring to regulate their study time. A classic paradigm suited to further explore this issue refers to the *allocation of study time*. Research on study time allocation observes how learners deploy their attention and effort. As already noted by Brown et al. (1983), the ability to attend selectively to relevant aspects of a problem solving task is a traditional index of learner's understanding of the task. Developmental studies on the allocation of study time examined whether schoolchildren and adults were more likely to spend

more time on less well-learned material (e.g., Masur et al. 1973; Dufresne and Kobasigawa 1989; Lockl and Schneider 2004). All of these studies reported an age-related improvement in the efficient allocation of study time. That is, older children (from age 10 on) spent more time studying hard items than they spent studying easy items, despite the fact that even many young children were able to distinguish between hard and easy pairs.

Thus, developmental differences were not so much observed in the metacognitive knowledge itself but in its efficient application to self-regulation strategies.

5 The Importance of Metacognition for Education

During the last three decades, several attempts have been made to apply metacognitive theory to educational settings (cf. Paris and Oka 1986; Moely et al. 1995; Palincsar 1986; Pressley 1995).

One interesting and effective approach to teaching knowledge about strategies was developed by Palincsar and Brown (1984). The "reciprocal teaching" procedure requires that teachers and students take turns executing reading strategies that are being taught with instruction occuring in true dialog. Strategic processes are made very overt, with plenty of exposure to modeling of strategies and opportunities to practice these techniques over the course of a number of lessons. The goal is that children discover the utility of reading strategies, and that teachers convey strategy-utility information as well as information about when and where to use particular strategies. Teachers using reciprocal instruction assume more responsibility for strategy implementation early in instruction, gradually transferring control over to the student (see Palincsar 1986, for an extensive description of the implementation of reciprocal instruction; see Rosenshine and Meister 1994, for a realistic appraisal of its benefits).

During the eighties and nineties of the last century, numerous studies explored the efficiency of strategy training approaches in school (for a review, see Schneider and Pressley 1997). The basic assumption was that although children in most cases do not efficiently monitor the effectiveness of strategies they are using, they can be trained to do so. For instance, in a training program carried out by Ghatala and colleagues (e.g., Ghatala et al. 1986) elementary school children were presented with paired-associate learning tasks. Before studying these lists, some children received a three-component training. They were taught (a) to assess their performance with different types of strategies, (b) to attribute differences in performance to use of different strategies, and (c) to use information gained from assessment and attribution to guide selection of the best strategy for a task. As a major result, it was shown that even children 7–8 years of age can be taught to monitor the relative efficacy of strategies that they are using and to use utility information gained from monitoring in making future strategy selections.

Another more large-scale approach concerns the implementation of comprehensive evaluation programs that aim at assessing the systematic instruction of metacognitive

knowledge in schools. As emphasized by Joyner and Kurtz-Costes (1997), both Moely and Pressley, with their colleagues, have conducted very ambitious programs of evaluating effective instruction in public school systems. For instance, Pressley and colleagues found that effective teachers regularly incorporated strategy instruction and metacognitive information about effective strategy selection and modification as a part of daily instruction. It seems important to note that strategy instruction was not carried out in isolation but integrated in the curriculum and taught as part of language arts, mathematics, science, and social studies. In accord with the assumption of the Good Information Processor Model outlined above (cf. Pressley et al. 1989), effective teachers did not emphasize the use of single strategies but taught the flexible use of a range of procedures that corresponded to subject matter, time constraints, and other task demands. On most occasions, strategy instruction occured in groups, with the teachers modeling appropriate strategy use. By comparison, the work by Moely and colleagues (e.g., Moely et al. 1995) illustrated that the effective teaching process described by Pressley and coworkers does not necessarily constitute the rule, and that effective teachers may represent a minority group in elementary school classrooms. Taken together, the careful documentations of instructional procedures carried out by Pressley, Moely, and their research groups have shown that there is a lot of potential for metacognitively guided instructional processes in children's everyday learning.

More recent research explores the utility of the metacognition concept in research with older children and adolescents, assessing the predictive potential of metacognitive knowledge and skillfulness in reading and math (e.g., Artelt et al. 2002; Veenman et al. 2005; see also the contributions in Desoete and Veenman 2006). Overall, these studies confirm the view that metacognitive knowledge and self-regulated, insightful use of learning strategies predicts math performance and reading comprehension in secondary school settings even after differences in intellectual abilities have been taken into account. They also give evidence that metacognitive knowledge relevant for school-related domains can still be effectively trained in late childhood and early adolescence.

References

Artelt C, Schiefele U, Schneider W (2002) Predictors of reading literacy. Eur J Psychol Educ 16:363–383

Astington JW, Jenkins JM (1999) A longitudinal study of the relation between language and theory of mind development. Dev Psychol 35:1311–1320

Borkowski JG, Milstead M, Hale C (1988) Components of children's metamemory: implications for strategy generalization. In: Weinert FE, Perlmutter M (eds) Memory development: universal changes and individual differences. Erlbaum, Hillsdale, NJ, pp 73–100

Brown AL, Bransford JD, Ferrara RA, Campione JC (1983) Learning, remembering, and understanding. In: Flavell JH, Markham EM (eds) Handbook of child psychology. Cognitive development, vol 3 Wiley, New York, pp 77–166

Cultice JC, Somerville SC, Wellman HM (1983) Preschoolers' memory monitoring: feeling-of-knowing judgments. Child Dev 54:1480–1486

DeLoache JS, Brown AL (1984) Where do I go next? Intelligent searching by very young children. Dev Psychol 20:37–44
Desoete A, Veenman M (eds) (2006) Metacognition in mathematics education. Nova Science, Haupauge, NY
Dufresne A, Kobasigawa A (1989) Children's spontaneous allocation of study time: differential and sufficient aspects. J Exp Child Psychol 47:274–296
Efklides A (2001) Metacognitive experiences in problem solving: metacognition, motivation, and self-regulation. In: Efklides A, Kuhl J, Sorrentino RM (eds) Trends and prospects in motivation research. Kluwer, Dordrecht, The Netherlands, pp 297–323
Flavell JH (1979) Metacognition and cognitive monitoring a new area of cognitive-developmental inquiry. Am Psychol 34:906–911
Flavell JH (2000) Development of children's knowledge about the mental world. Int J Behav Dev 24:15–23
Flavell JH, Wellman HM (1977) Metamemory. In: Kail R, Hagen J (eds) Perspectives on the development of memory and cognition. Erlbaum, Hillsdale, NJ, pp 3–33
Flavell JH, Miller PH, Miller SA (2002) Cognitive development, 4th edn. Prentice-Hall, Inc., Englewood Cliffs, NJ
Garner R (1987) Metacognition and reading comprehension. Ablex, Norwood, NJ
Ghatala ES, Levin JR, Pressley M, Goodwin D (1986) A componential analysis of effects of derived and supplied strategy-utility information on children's strategy selections. J Exp Child Psychol 41:76–92
Goswami U (2008) Cognitive development – The learning brain, 2nd edn. Psychology Press, Hove, UK
Hasselhorn M (1990) The emergence of strategic knowledge activation in categorical clustering during retrieval. J Exp Child Psychol 50:59–80
Joyner MH, Kurtz-Costes B (1997) Metamemory development. In: Cowan N (ed) The development of memory in childhood. Psychology Press, Hove East Sussex, UK, pp 275–300
Justice EM (1985) Preschoolers' knowledge and use of behaviors varying in strategic effectiveness. Merrill Palmer Q 35:363–377
Koriat A (1993) How do we know that we know? The accessibility model of the feeling of knowing. Psychol Rev 100:609–639
Kuhn D (1999) Metacognitive development. In: Balter L, Tamis-LeMonda CS (eds) Child psychology: a handbook of contemporary issues. Psychology Press, Philadelphia, PA, pp 259–286
Kuhn D (2000) Theory of mind, metacognition, and reasoning: a life-span perspective. In: Mitchell P, Riggs KJ (eds) Children's reasoning and the mind. Psychology Press, Hove, UK, pp 301–326
Lockl K, Schneider W (2002) Developmental trends in children's feeling-of-knowing judgements. Int J Behav Dev 26:327–333
Lockl K, Schneider W (2004) The effects of incentives and instructions on children's allocation of study time. Eur J Dev Psychol 1:153–169
Lockl K, Schneider W (2006) Precursors of metamemory in young children: the role of theory of mind and metacognitive vocabulary. Metacognition Learn 1:15–31
Lockl K, Schneider W (2007) Knowledge about the mind: links between theory of mind and later metamemory. Child Dev 78:148–167
Masur EF, McIntyre CW, Flavell JH (1973) Developmental changes in apportionment of study time among items in a multitrial free recall task. J Exp Child Psychol 15:237–246
Metcalfe J (2002) Is study time allocated selectively to a region of proximal learning? J Exp Psychol Gen 131:349–363
Moely BE, Santulli KA, Obach MS (1995) Strategy instruction, metacognition, and motivation in the elementary school classroom. In: Weinert FE, Schneider W (eds) Memory performance and competencies: issues in growth and development. Erlbaum, Mahwah, NJ, pp 301–321
Nelson TO (1996) Consciousness and metacognition. Am Psychol 51:102–116

Nelson TO, Narens L (1990) Metamemory: a theoretical framework and new findings. In: Bower G (ed) The psychology of learning and motivation: advances in research and theory, vol 26. Academic Press, New York, pp 125–173
Nelson TO, Narens L (1994) Why investigate metacognition? In: Metcalfe J, Shimamura AP (eds) Metacognition. Knowing about knowing. MIT Press, Cambridge, MA, pp 1–25
Palincsar AS (1986) The role of dialogue in providing scaffolded instruction. Educ Psychol 21:73–98
Palincsar AS, Brown AL (1984) Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognit Instruct 1:117–175
Paris SG, Oka ER (1986) Children's reading strategies, metacognition, and motivation. Dev Rev 6:25–56
Perner J (1991) Understanding the representational mind. MIT Press, Cambridge, MA
Pressley M (1995) What is intellectual development about in the 1990s? In: Weinert FE, Schneider W (eds) Memory performance and competencies: issues in growth and development. Erlbaum, Hillsdale, NJ, pp 1–25
Pressley M, Afflerbach P (1995) Verbal protocols of reading: the nature of constructively responsive reading. Erlbaum, Hillsdale, NJ
Pressley M, Borkowski JG, Schneider W (1989) Good information processing: what it is and what education can do to promote it. Int J Educ Res 13:857–867
Roebers C, von der Linden N, Howie P, Schneider W (2007) Children's metamemorial judgments in an event recall task. J Exp Child Psychol 97:117–137
Rosenshine B, Meister C (1994) Reciprocal teaching: a review of the research. Rev Educ Res 64:479–530
Ruffman T, Slade L, Crowe E (2002) The relation between children's and mother's mental state language and theory-of-mind understanding. Child Dev 73:734–751
Schneider W (1986) The role of conceptual knowledge and metamemory in the development of organizational processes in memory. J Exp Child Psychol 42:218–236
Schneider W, Lockl K (2002) The development of metacognitive knowledge in children and adolescents. In: Perfect TJ, Schwartz BL (eds) Applied metacognition. Cambridge University Press, Cambridge, UK, pp 224–257
Schneider W, Lockl K (2008) Procedural metacognition in children: evidence for developmental trends. In: Dunlosky J, Bjork RA (eds) A handbook of metamemory and memory. Erlbaum, Mahwah, NJ
Schneider W, Pressley M (1997) Memory development between 2 and 20. Erlbaum, Hillsdale, NJ
Schneider W, Perner J, Bullock M, Stefanek J, Ziegler A (1999) Development of intelligence and thinking. In: Weinert FE, Schneider W (eds) Individual development from 3 to 12: findings from the munich longitudinal study. Cambridge University Press, Cambridge, MA, pp 9–28
Schneider W, Hünnerkopf M, Kron-Sperl V (2009) The development of young children's memory strategies: evidence from the Würzburg Longitudinal Study. European Journal of Developmental Psychology 6:70–99
Schunk DH, Zimmerman BJ (eds) (1998) Self-regulated learning: from teaching to self-reflective practice. Guilford, New York
Shimamura AP (2000) Toward a cognitive neuroscience of metacognition. Conscious Cogn 9:313–323
Sodian B (2005) Theory of mind. the case for conceptual development. In: Schneider W, Schumann-Hengsteler R, Sodian B (eds) Interrelationships among working memory, theory of mind, and executive function. Erlbaum, Mahwah, NJ, pp 95–130
Sodian B, Schneider W, Perlmutter M (1986) Recall, clustering, and metamemory in young children. J Exp Child Psychol 41:395–410
Veenman M, Kok R, Blöte A (2005) The relation between intellectual and metacognitive skills in early adolescence. Instr Sci 33:193–211
Wellman HM (1977) Preschoolers' understanding of memory-relevant variables. Child Dev 48:1720–1723
Wellman HM (1985) A child's theory of mind: the development of conceptions of cognition. In: Yussen SR (ed) The growth of reflection in children. Academic Press, New York, pp 169–206
Zabrucky K, Ratner HH (1986) Children's comprehension monitoring and recall of inconsistent stories. Child Dev 57:1401–1418

Understanding Apes to Understand Humans: The Case of Object–Object Relations

Josep Call

Abstract Animal cognition has grown exponentially in the last decade and more than ever has established close links with human cognition to jointly explore the mechanisms, ultimate functions, and evolution of cognition. The comparative method plays a key role in this endeavor. Knowledge about object–object relations is a good example of this growth, but just like the rest of animal cognition, it has been dominated by a two-tiered framework (perceptually based vs. conceptually based). Much of animal cognition is routinely reduced to associations between stimuli and responses. I argue that this view is too narrow with regard to apes' knowledge about object–object relations. Instead, I propose that apes distinguish between arbitrary and causal relations between objects. This means that apes not only associate the presence of certain stimuli with certain events but also attribute a causal role between the presence of those objects and certain events.

1 Introduction

Humans cannot fly, we can barely swim and we can be easily outsprinted but apparently not outrun (see Bramble and Lieberman 2004) by a number of animals, several of which would not hesitate before having us for a meal. Our poor athletic skills are matched by our unremarkable body features. Except for our bipedal stance, we lack striking morphological specializations such as a peacock's tail, a giraffe's neck or an elephant's trunk. However, there is one thing we can be proud of: our brain. Although at a first glance it may not seem particularly remarkable – it is not the largest (or most convoluted) in the animal kingdom and it does not possess any distinctive features that set it qualitatively apart from other species – we have a much larger brain compared with what would correspond for an animal of our size.

J. Call
Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig, D-04103, Germany
e-mail: call@eva.mpg.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_15, © Springer-Verlag Berlin Heidelberg 2010

Our brain indeed is the seat of what we consider the thing that in reality sets us apart from other species – our intelligence.

There is no question that humans are intelligent creatures. We are capable of engaging in a number of higher mental processes as this volume clearly shows. The main source of this vast amount of knowledge about humans is logically research on humans both in their everyday lives and in the laboratory. We can study how humans routinely solve complex problems and what is the impact of certain forms of instruction and enculturation on cognitive processes. There is a second avenue to learn about human cognition: Research on animal cognition. Although studying nonhumans to learn about humans may seem a paradoxical proposal at a first glance, I will argue that these two complementary approaches have much to offer to each other.

I will go as far as saying that for some questions about human cognition, animal cognition holds the key answers. Thus, I will devote the first section of this chapter to briefly explore three ways in which animal cognition can be informative. Next, I will sketch, briefly again, three frameworks that have shaped the kinds of questions that scientists ask about animal cognition, particularly when comparing animal and human cognition. In the remaining of the chapter, I will illustrate how animal cognition can establish connections with topics that have been traditionally been studied in humans. Thus, this section is devoted to explore the question of object–object relations and causal knowledge in inferential and tool-using tasks in the great apes using some of the latest work from our laboratory on this topic. Finally, I will close with some considerations regarding the nature of nonhuman knowledge about the world.

2 Why Study Animal Cognition?

I think that it was the primatologist Irven deVore, who once remarked that if NASA were to discover a chimpanzee-like creature living in Mars, the agency would spend billions of dollars to study it. My guess is that the reasons for engaging in such enterprise would fall into one of two broad categories. First of all, we would study the creature for the sake of advancing our knowledge on that particular organism – an enterprise that human cultures have been practicing since antiquity on a number of subjects ranging from astronomy to natural history. In fact, this drive to advance our knowledge is one main reason why we study animal cognition today. We want to know how animals acquire, store, and manipulate information to solve problems related to their survival and reproduction.

The second reason for studying the Martian creature would be to learn something about ourselves in the process. Although this may seem a quite paradoxical reason at first, and some may argue that the best way to learn about humans is to study ourselves, this may be too narrow of a view. Indeed, one key analytical tool for many subjects is to compare entities, be it atomic particles, brains, languages, whole organisms, ecosystems, or galaxies. Why are comparisons so important?

Because they are a contrasting device that highlight, in fact are essential, to uncover the similarities and differences between entities. Without knowing what other animals can and cannot do, it is unclear that we can pinpoint what makes us human. Just as an example, human tool-using and tool-making skills had been heralded as the human rubicon – material culture was what made us humans. This idea, which was very popular in the first half of the twentieth century – was abandoned in the 1960s when it was discovered that chimpanzees regularly made and used tools in the wild.

Comparisons are also an invaluable heuristic tool to discover the significant things that a species does not do. A detailed observation can reveal the things that a species does, for instance, it uses tools, vocalizes, and feeds on termites. However, it cannot inform us about what the species does not do given that there is an infinite amount of things that a species does not do and we have no way of knowing whether those missing things are significant or not. Unless, of course, we have another species with which to compare. Only then one can see the significant things that were missing.

Finally, comparisons between species are crucial to make inferences about cognitive evolution. Chimpanzees or bonobos are not our ancestors, in fact none of the living primates is, but comparisons between humans and apes and other animals offer us important clues about what may have changed and what may have not in human evolution. For instance, humans unlike any other primate can produce novel vocalizations. In contrast, all great apes can use and make tools quite easily. Therefore, we can postulate that whereas voluntary vocal production evolved quite recently after our ancestors split from the chimpanzee-bonobo ancestor tool-use evolved much earlier and it was already present in the common ancestor of all extant great apes. Other species, such as parrots and New Caledonian crows, can control vocalizations and fashion and use tools, respectively. However, the mechanism is different when compared to humans.

In sum, we study animals for three main reasons: (1) to advance our knowledge on particular species, (2) as a contrasting device that allows us to gauge the importance of similarities and differences and to “uncover” hidden traits, and (3) as a tool for making inferences about cognitive evolution.

3 Comparing Human and Animal Cognition

Theoretical frameworks guide, to a large extent, the kinds of hypothesis that we pose and how we examine them empirically. The study of animal cognition, especially when applied to comparing human and animal cognition, has been traditionally dominated by a dichotomous framework of reference based on considering only two extreme positions (see Tomasello and Call 1997; Penn and Povinelli 2007). Such an approach has deep roots in the Cartesian philosophical tradition. In effect, comparisons between human and animal cognition have generally contrasted low-level mechanisms (typically attributed to animals) with high-level mechanisms

(typically attributed to humans). For instance, the last decade has seen a lively discussion about whether animals can attribute mental states to others. Proponents of the so-called low-level mechanism position argue that animals react to observable stimuli whereas the proponents of the so-called high-level position argue that animals may engage in metarepresentational thought. The usual outcome of this exercise is that neither of the two positions is fully supported by the data. We have argued that a third alternative that lies between them is the one that typically best explains the data (Tomasello and Call 1997; Call and Tomasello 2005a, 2008) – not that the explanation is complete but it produces the best fit. In the case of mental state attribution, this option postulates that animals go beyond observable behavior and infer psychological states although they may not be as cognitively sophisticated as some forms seen in humans.

However, obtaining a better data fit is not the only reason for exploring the existence of alternatives to the so-called low- and high-level mechanisms. There are at least two other reasons to diversify the search space. The first reason is to have a larger set of pieces with which to attempt to reconstruct both the ontogeny and the phylogeny of cognition. It seems that the so-called low- and high-level mechanisms are too far apart to attempt such a reconstruction. Nevertheless, it is conceivable that low-level mechanisms are instrumental in fostering the emergence of other mechanisms which in turn can lead to the emergence of the so-called high-level mechanisms.

The second reason for seeking a more diverse set of pieces is that although humans use certain high-level mechanisms to solve problems, they do not use them all the time. Instead humans often rely upon computationally less-demanding habits and rules of thumb to solve problems. Nevertheless, humans can reengage high-level processes when the mechanisms based on rules of thumb do not produce the desired effects. This means that humans, and possibly other animals, have at their disposal a variety of mechanisms to solve the same task, not just one as the two opposing positions may suggest. Coexistence of this sort is not a strange notion if one considers that redundancy is a major hallmark of how our brains are built and how they work. Once one accepts that multiple cognitive mechanisms coexist, this opens up the possibility for dynamical interactions between them. According to this more dynamic approach, subjects' responses result from the interaction and contribution of multiple mechanisms including epistemic aspects, perceptual and motor biases and constraints. Moreover, responses do not occur independently of the context, the context or the testing conditions are an essential part of the how cognition is deployed. One major task for future research will be to map the various mechanisms and how they interact with each other to produce effective responses.

In sum, comparative cognition has been traditionally dominated by the two-tiered approach based on contrasting a low- and a high-level mechanism. Such characterization is too simplistic and often requires a serious reevaluation. Once multiple mechanisms are considered, the door is open to a greater interaction between mechanisms understood as information processing devices. In the remaining of the chapter, I will present some evidence about causal reasoning in the great apes. I think that this evidence, just like the one on mental state attribution illustrates the third alternative – apes do not engage in the sophisticated forms of causal reasoning

observed in humans but they are not either easily accounted for by mechanisms based on the acquisition of certain heuristics. Moreover, the interaction between various systems will be illustrated in the last part of the next section (motor and knowledge components).

4 Object–Object Relations and Causal Knowledge

Causal inference is one of the complex cognitive processes that has been extensively investigated in humans (see papers in Sperber et al. 1995; Gopnik and Schulz 2007; this volume). Moreover, some forms of causal knowledge such as counterfactual reasoning or postulating unobservable constructs (e.g., gravity) are thought to be uniquely human (Povinelli 2000). It is reasonably safe assumption that animals lack the analytic sophistication that some humans can display in disciplines such as logic or physics. Yet, does that lack of sophistication mean that animals possess no causal reasoning abilities? Are they restricted to innate predispositions and trial-and-error learning? This is the question that will occupy us for the remaining of the chapter. However, before presenting the evidence on causal inference in great apes allow me to provide some clarification about the meaning of causal inference as it is used here. This is necessary because this term may be understood in different ways depending on the orientation of the scholars that study them.

Here, I use causal knowledge to refer to information about object–object interactions and relations that subjects have encoded and stored in their memories. For Piaget, knowledge about interactions between objects, not just knowledge about actions on the self or on objects, was one of the first manifestations of causality. Inference is the mechanism that allows subjects to use and combine old and new information to solve problems that they have not experienced before. Therefore, causal inferences consist of inferences made on the basis of information about object–object relations. Solutions based on this mechanism typically appear suddenly without trial-and-error learning. In the next two sections, we explore the use of object–object relations to infer the location of hidden objects and to use tools to bring rewards within reach while avoiding obstacles.

4.1 Inferential Reasoning

The inferential abilities of primates have been documented in various domains (see Call and Tomasello 2005b; Tomasello and Call 1997 for reviews). Monkeys and apes can infer the location of hidden objects based on their past association with certain landmarks, the geometric disposition of other target objects, or the successive displacements behind several barriers (Menzel 1996; Hemmi and Menzel 1995; Call 2001). There are also studies on transitive inference (Boysen et al. 1993; Gillan 1981; Yamamoto and Asano 1995) and various studies that have found

evidence of inference by exclusion (Premack and Premack 1994; Hashiya and Kojima 2001; Tomonaga 1993; Call 2006a; Call and Carpenter 2001). In these studies, the information that subjects exploit to designate the correct alternative is based either on arbitrary associations between stimuli (e.g., transitive inference) or the initial location and potential trajectories of the target object(s) (e.g., object permanence).

Much less is known about inferences based on object–object interactions other than chimpanzees and rhesus macaques associate certain object transformations with particular outcomes (e.g., cut apple with knife, Premack and Premack 1983; Hauser and Spaulding 2007). Investigating the use of object–object interactions for inferential purposes is particularly interesting because animals experience a variety of object–object relations in their everyday lives. Taking them into account and extracting information from them are two important aspects that contribute to the successful adaptation of the individual to its environment. Subjects experience that the shape or the color of a fruit indicates its ripeness, observe that fruits that are not attached to a branch fall to the ground, and once they reach the ground they continue to exist despite having momentarily disappeared from sight. Although those cases exemplify certain object properties and relations between objects, they have very different status. I distinguish two main types of relations between objects: causal and arbitrary (Call 2006b). Let me provide an experimental example on arrested motion to illustrate this distinction.

We presented an ape with a table and showed her a piece of monkey chow. After raising an opaque screen between the subject and the table, we deposited the reward on the table so that the subject could not see it anymore and covered it with a small rigid board that acquired a slanted orientation *caused* by the presence of the reward underneath. The orientation of the board was such that it blocked the subject's visual access to the reward. We also placed a second identical board next to the slanted one but flat on the table since there was no reward under it (see Fig. 1). Upon removal of the opaque screen that was blocking the ape's view of the table, the ape could see two small boards, one slanted and the other flat, but no reward in sight. The ape's task consists of finding the reward on the first attempt.

In a second condition, this time instantiating an arbitrary relation, we show the subject that the table has two receptacles where the reward can be introduced. Behind the opaque screen, we proceed to place the reward inside one of these two receptacles and cover both holes. The empty hole is covered with a flat board whereas the baited hole is covered with a wedge that when placed on the table appears like the slanted wedge of the causal condition. Again, we remove the screen and subjects can select one of the two alternatives. Apes presented with this task select the baited alternative on the first trial in the causal but not the arbitrary condition (Call 2007a). This means that they can distinguish the two alternatives even though perceptually they are very similar.

Two other pieces of information are important. First, although subjects selected the baited alternative at chance levels in the arbitrary condition, providing a couple dozens of additional trials did not help them improve their performance. This is not surprising if one takes into account that learning arbitrary relations is hard for nonhuman apes (Call 2006b). There is the widespread belief that apes are very fast

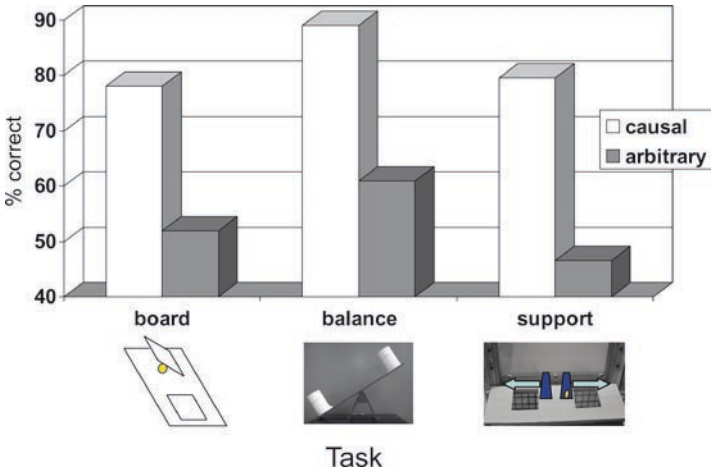

Fig. 1 Percent of correct trials in three tasks as a function of the nature of the relations between the presence of the reward and the elements of the task

at learning arbitrary relations including some complex relations. This is demonstrably false. They can be trained to solve arbitrary problems quickly (as in learning set formation, Harlow 1949), but they do not initially solve them easily.

Second, if subjects that are solving the causal condition are presented with the arbitrary condition, their performance decreases dramatically. This result is surprising from the point of view of stimulus generalization because those subjects that have solved the slanted board should continue picking the wedge since it is perceptually very similar to the slanted board. But this is not what they do. Many subjects start choosing randomly. Thus, subjects are able to infer the food location in the causal condition but they are not able to learn it in the arbitrary condition. Obviously, if given enough trials they would end up learning but here the mechanism may be very different. Our hypothesis is that they are able to solve the causal task in the first trial because they infer that the reward is causing the slanted orientation.

4.2 Two Additional Domains: Weight and Support

Although the results of this experiment were clear, one may wonder whether this pattern of results is peculiar to this setup. In other words, can apes make the causal vs. arbitrary distinction in other tasks? Let us examine two other studies aimed at the same question but with different arrangements. One investigates whether subjects can infer the location of food based on its weight, or better the effect that its weight has on other objects, while the other investigated whether subjects know that unsupported objects invariably fall.

Chimpanzees are sensitive to the effect that their own weight has on pliable vegetation (otherwise, they would fall from trees all the time), yet much less is

known whether they can use object weight to make inferences about the location of rewards. Hanus and Call (2008) presented chimpanzees with two opaque cups mounted on opposite sides of balancing beam kept in equilibrium by a pivot located under its center of gravity. In one condition, the experimenter hid a reward inside one of the cups and released the beam which resulted in the baited cup moving downwards and the empty cup moving upwards. Once the beam had reached this new equilibrium, subjects were allowed to select one of the cups. Obviously, the correct cup was always the lower cup. We compared the causal condition with two control conditions. In the static control condition we assessed whether subjects preferred to select the lower cup rather than the upper cup when they were mounted on a static inclined beam so that the weight of the reward was not responsible for the final orientation of the cups. In the external cause condition, the setup was identical to the causal condition except that when the experimenter released the beam after baiting, that maintained its equilibrium and the experimenter pushed physically down the beam after it reached its final slanted orientation. Thus, the experimenter, not the weight of the reward was responsible for the change in orientation. Chimpanzees selected the baited cup in the causal condition but not in any of the control conditions (see Fig. 1). Moreover, subjects that were performing at above chance levels in the experimental condition responded at chance levels upon receiving the static control condition. Conversely, subjects that were responded at chance levels in the static control condition began responding above chance in the causal condition. Equally remarkable is the difference between that causal and the external cause condition given that the information about the beam's displacement and the final position of the baited cup were identical in both conditions.

Let us turn now our attention to a task about object support. Again, the idea is the same as the two previous tasks – can subjects make inferences about the location of a reward based on the effects that it has on other objects. Martin-Ordas and Call (in press) presented apes with a platform that had two square holes cut on it so that it created three solid areas on the front part of the platform; one central area and two smaller areas next to each hole on each side of the platform. One hole was covered with a transparent piece of plastic and the other hole was left uncovered. Two opaque plastic cups are placed upside down side by side on the central area of the platform next to the holes. The experimenter showed a reward to the subject and behind a screen placed it under one of the cups so that the subject did not see its final destination. After the baiting was completed, the experimenter removed the screen and laterally displaced each cup from the central area to the side so that each cup crossed over the hole closest to them. After both cup displacements were completed, the ape could select one of the cups by touching it. In order to avoid the noise that the reward would make when it fell through the open hole, we never displaced the reward over the open hole but it was always displaced over the covered hole.

We found that apes selected the baited cup above chance levels both overall and in the first trial butfailed to do so if both holes were covered with opaque or transparent pieces of plastic (Fig. 1). This ruled out the possibility that subjects used inadvertent cues left by the reward or the experimenter to solve the problem. Apes also failed to select the baited cup if the displacements occurred when both holes were covered, but later, one hole was covered with a transparent piece of plastic and the other was

left uncovered. Since this is the same perceptual configuration that subjects encountered at the time of choice in the experimental condition, we can rule out that subjects had a predisposition for avoiding uncovered holes regardless of the reward displacements. It also indicates that subjects were not choosing based on the final configuration alone (i.e., simply avoiding cups next to the hole).

Once again, the two key findings are confirmed and strengthened. First, apes make a distinction between causal and arbitrary relations between stimuli. Furthermore, the causal–arbitrary distinction is a robust phenomenon found in various situations. Second, there is no evidence of learning to solve the problem via conditional discrimination – again with enough trials they could but without such training they engage an inferential rather than associative mechanism. Additionally, these studies also help us rule out the possibility that subjects have a predisposition to respond to certain stimuli in certain ways. For instance, the balance control has the same movement and position of the causal balance condition. Several of the control conditions in the trap task showed that subjects did not have a predisposition to avoid holes.

This last study also illustrates one interesting points compared to the other two previous studies. One has to do with the type of information that can be used to make inferences. Subjects were able to locate the reward even though displacing both cups laterally produced no observable effect. Recall that in the slanted board task and the balance task the reward caused an observable effect on other objects. Since control tests showed that subjects did not avoid the cup next to the gap (or preferred the cup next to the blocked hole), we hypothesized that subjects used the lack of a falling reward when the cup crossed the open gap as an indication that the reward was not there. In other words, subjects used information about something that did not happen to infer the location of the reward. This kind of inference has been described in great apes before in the context of finding the food located in one of the two cups by the noise the food makes when the cup that contains the food is shaken. In particular, subjects selected a cup that was lifted but not shaken compared to a cup that was shaken even though neither cup produced any sound but only the lifted cup could contain the food, because otherwise the shaken cup would have produced the noise. Penn and Povinelli (2007) have criticized this interpretation by arguing that the result may have arisen from the previous experience. Although the data do not support the interpretation that subjects learned to respond the way they did during the test (see also Herrmann et al. 2007), it is possible that they had learned to respond appropriately before the task. Allow me to defer the discussion of this issue for the final section of the chapter and let me instead use the gap task to make a connection with the literature on tool-use, particularly tool-use as a way to get out-of-reach rewards while avoiding obstacles such as traps.

4.3 Tool-Use

Traditionally, much of the work on causal knowledge in primates has focused on tool-use. Researchers have used relatively simple tasks such as the support or the stick problem in which subjects have to retrieve an out-of-reach reward by using a

tool (Natale et al. 1988; Piaget 1952; Spinozzi and Potì 1989). Even more complex yet, are those tasks in which subjects have to overcome some obstacle (not just the distance) on the way to getting the reward with a tool. One of the most well-known tasks of this kind is the trap-tube (Visalberghi and Limongelli 1994; Limongelli et al. 1995). Subjects are faced with a transparent tube that has a trap in its center and a reward placed inside the tube next to the trap and outside of the subject's direct reach and they can use a stick whose diameter is slightly smaller than the inner diameter of the tube. The solution to this problem consists of inserting the stick inside the tube and pushing the reward away from the trap and outside of the tube. This task has proven extremely difficult to solve for capuchin monkeys, chimpanzees and birds as only a minority of subjects solved this task even after dozens of trials (Tebbich and Bshary 2004) (Fig. 2).

Some variations on the trap tube aimed at simplifying the task have produced similar negative results (Call, in press, a review). For instance, Povinelli (2000) presented chimpanzees with a pair of rakes each with a reward in front of them. Crucially, one rake also had a trap in front of it while the other simply had a painted patch of the same dimensions of the trap (i.e., fake trap). In order to succeed, the only thing that subjects had to do was to pull the rake placed behind the fake trap since pulling the other rake invariable sent the reward into the trap. All subjects except one failed to solve the task, and even this subject did not pass subsequent control tasks –something that has also been observed in the original trap-tube task (Limongelli et al. 1995; Povinelli 2000). This has led several authors to conclude that subjects may have used a perceptual strategy based on using the position of the trap to determine the appropriate insertion point but without understanding that the position of the reward with respect to the trap hole is the critical feature in this task. Povinelli (2000) concluded that apes had a limited understanding of the physical properties of the trap. Subsequent studies carried out with capuchins and gibbons using a similar paradigm, also concluded that subjects did not have a total comprehension of the elements of the problem but that subjects might have learned certain associative rules to solve the problem (Fujita et al. 2003; Cunningham et al. 2006).

In recent years, however, researchers have began to find more positive results on trap tasks provided certain modifications are implemented. In particular, a performance improvement has been observed when subjects can choose what actions to use (pulling or pushing) (Martin-Ordas et al. 2008; Mulcahy and Call 2006), where to insert the tool (Girndt et al. 2008), or the need to use a tool is completely eliminated

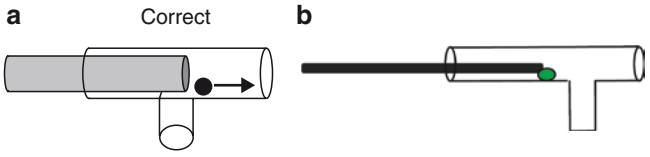

Fig. 2 Two versions of the trap-tube task: (**a**) Visalberghi and Limongelli 1994, (**b**) Mulcahy and Call 2006. Note the differences in the diameter of the tool and the position of the trap

(Seed et al. 2009). This last result is particularly illuminating because subjects who passed a version of the task that required no tools were incapable of transferring to a task that required tools, whereas those who had mastered the task with tools had no difficulty transferring to non- tool-using version of the task. Thus, the difficulties with solving the task may have more to do with the use of tools than an understanding of the properties of the traps. This means that certain task requirements may mask the knowledge that subject possess about the effect that traps may have onrewards. Yet, there is some evidence suggesting that apes treat traps in similar ways even if they differ in their shape, size, and location as well as the response required to get the reward. In particular, Martin-Ordas and Call (in press) found some positive correlations in apes' performance on some trap tasks including the gap task, which as previously indicated requires no tools.

5 The Nature of Causal and Arbitrary Relations

We have already alluded to the distinction between causal and arbitrary object–object relations in previous sections. Let me return to it by elaborating further on some aspects of this distinction. One thing that all examples about causal object relations have in common is that they are invariably anchored in physical laws that govern object–object interactions on planet Earth. Objects continue to exist even though we cannot see them, they fall to the ground if they are unsupported and they can potentially affect the orientation and movement of other objects that they collide with. In contrast, arbitrary relations lack such anchorage in the laws of physics. There is no physical (causal) law that determines that a certain color or tone be associated with the presence or the absence of a reward. Interestingly, note that much of the learning literature has used arbitrary relations between stimuli, for instance associating a green light with reward delivery and a red light with no reward. This choice is partly understandable to avoid the thorny issue of predisposition in learning, but it may have had the undesirable effect of underestimating the importance of natural causal relations. This is not to say that associating stimuli to create arbitrary relations is unimportant, but it is not the only way to solve problems, a point already expressed by Köhler (1925) when referring to Thorndike's work on puzzle boxes.

Causal and arbitrary relations can have the same predictive power, but only causal relations have in addition explanatory power. An organism capable of operating at the level of arbitrary relations can predict that food will be found under the slanted board or that the lower of two cups mounted on a beam is the one holding the food. However, only an organism capable of operating at the level of causal relations will, in addition, be able to understand that the presence of the food caused the observed changes. This may also be the reason by tasks based on causal relations are solved in the first trial and arbitrary relations are not. While the latter entails *learning* to find the food location, the former entails *predicting* the food location based on the available information. This means that subjects have the possibility to

solve the causal tasks in the first trial before they receive any feedback. I propose that the great apes, at least, can extract the richer information contained in causal events by engaging mechanisms of causal (not just correlational) analysis.

5.1 On the Epigenesis of Causal Relations

One could argue that the distinction between causal and arbitrary relations is not as deep as I am proposing. The reason subjects perform better with problems based on causal rather than arbitrary relations is because they are predisposed to attend to the former and not the latter. In fact, given that subjects will invariably encounter causal relations between events in their everyday lives, it would not be surprising to find that certain animals are predisposed to pick up certain types of relations, in the same way that they are predisposed to perceive light at a certain wavelength or sound at a particular frequency. However, this does not mean that all possible responses are preprogrammed. There is a difference between having a predisposition to pick up certain relations and using those predispositions to construct a set of cognitive structures. If humans are any indication, cognitive development results from the interaction between innate predispositions and the inputs from the outside world (Baillargeon 1995; Karmiloff-Smith 1992).

There is, however, another variety of predisposition – acquired rather than innate predispositions. Subjects may respond better to tasks instantiating causal rather than arbitrary relations because they have encountered similar stimuli before the test took place. Penn and Povinelli (2007) propose that this is the explanation for the apes' successful performance in the inferential task based on auditory cues mentioned earlier. In particular, apes may have encountered containers filled with food in the past that when shaken made noise and others that remained silent and were empty. It is conceivable that they could have learned an association between the production of the auditory cue and the presence of food. The same logic could potentially explain other more recent studies presented in the current chapter. But, is this a viable explanation?

There is no question that experience with environmental stimuli plays a key role in subjects responding appropriately to the tests that we present them. In fact, without experiencing events in the environment subjects would not be able to solve our tests. However, experiences do not simply accumulate (as in a library) but they are the raw material that combined with predispositions and processed allows individuals to acquire a new cognitive organization. Iriki and colleagues have documented the substantial neurological changes that occur in macaques trained to use tools (Iriki and Sakura 2008). Therefore, the key question is not whether experience plays a role but what type of information subjects extract from the environment and how they process it to solve this task. One trivial possibility is that subjects have experienced the same problem that we administered them. We know that this is not the case because those subjects did not have access to the test stimuli prior to the test, and in fact, they had not been tested on those tasks before.

Another possibility is that although subjects had not experienced an identical task, they had experienced similar tasks and they were able to generalize from those past experiences to the testing situation. Thus, the presence of particular stimuli (e.g., noise, cup moving down, object with slanted surface) could be used to predict the location of food. Although this possibility represents a greater level of abstraction than facing an identical task solved in the past, we can also rule out this possibility because subjects repeatedly failed tests instantiating arbitrary relations but passed their causal counterparts even though both tests shared the same key features. At the very least, this means that subjects can detect the difference between these conditions but the difference, I argue, lies not on the features of the objects but on the nature of the object–object relations. In fact, the truly surprising thing is that apes that solved a task instantiating causal relations from the first trial on were unable to continue performing well as soon as causal relations were replaced by arbitrary relations.

This result is at odds with what one would predict if all that subjects were doing was associating certain features with certain outcomes and generalizing from those experiences to nearly identical exemplars. In reality, this is not what the apes do. Elsewhere, I have argued that quite often researchers are willing to invoke associative processes and stimulus generalization as the explanation for many phenomena both in physical and social cognition (Call 2003, 2007b). However, those hypotheses are rarely tested and, when they are, one often finds that they do not explain the data satisfactorily (Call 2007b). I will repeat again that there is ample evidence that individuals can use arbitrary relations between stimuli to solve problems, but at the same time it is important to emphasize that arbitrary relations, at least for the great apes, do not appear to cover the whole field. Hypotheses put forward to explain a particular phenomenon require empirical verification. Otherwise, it is easy to fall into the traditional two-tiered conjecture that animals solve problems using so-called low-level mechanisms and humans solve problems using so-called higher level mechanisms.

I would like to finish with a thought experiment. Apes solved problems that involved causal relations in the first exposure to the problem before they received any feedback. In contrast, they did not learn to solve nearly identical tasks in which the relation between stimuli was arbitrary. However, given enough trials, subjects would have learned to respond equally well to the task based on arbitrary relations. Thus, after sufficient exposure, arbitrary and causal relations at least at the level of performance can become nearly indistinguishable. The key question is whether such similarity is real or apparent. In other words, if an arbitrary relation were treated as a logical necessity, would subjects come to view it with the same status as a physical law? Let me offer one example to clarify this point. We know that unsupported objects fall. We have seen examples many times and we experience a strong uneasiness upon witnessing a violation of this law (e.g., a book floating in mid-air) and a great sense of relief when we can explain the violation (e.g., oh, it is an image from the International Space Station). Let us suppose that we could arrange that a child would invariably find objects every time he opens a blue box. All other boxes are just our normal boxes – sometimes there is something inside,

sometimes there is not. Would that child attribute box "blueness" the same status as gravity? More importantly, would blueness be perceived as the cause for the box not being empty?

6 Conclusion and Future Directions

Animal cognition has shown an exponential growth in the last decade. It is safe to say that more than ever, animal and human cognition have established links to jointly explore the causal mechanism, the ultimate function and the evolution and ontogeny of cognition. However, these links need to be further strengthened and diversified. There is much to be gained by doing so.

Causal inferences is an area that has grown steadily over the years but it has been dominated, just like the rest of animal cognition, by a two-tiered framework (perceptually based vs. conceptually based). I have argued that the knowledge that apes have about object–object interactions cannot be reduced to associations between stimuli and responses. It is much richer than that and I have presented evidence suggesting that apes, at least, distinguish between arbitrary and causal relations between objects. Such a distinction allows apes to not just expect the presence of certain objects but also to infer and perhaps even understand to some level that some objects affect others and that the presence (or absence) of some objects causes the observed effects on other objects.

Two features of this framework seem particularly promising for developing the field further in the next few years. First, some attention should be devoted to the interaction between the various types of information that individuals possess. Even the same stimuli can generate different types of information depending on the processing it receives. Second, the inter-individual variability in the use and the genesis of the three types of information requires a more in-depth treatment. All too often inter-individual variability is ignored as statistical noise when in reality it could explain the inter-individual variability that exists in problem solving both between and within species.

Acknowledgments I thank Amanda Seed for her helpful comments on a previous version of this manuscript.

References

Baillargeon R (1995) Physical reasoning in infancy. In: Gazzaniga MS (ed) The cognitive neurosciences. MIT Press, Cambridge, MA, pp 181–204

Boysen ST, Berntson GG, Shreyer TA, Quigley KS (1993) Processing of ordinality and transitivity by chimpanzees (*Pan troglodytes*). J Comp Psychol 107:208–215

Bramble DM, Lieberman DE (2004) Endurance running and the evolution of *Homo*. Nature 432:345–352

Call J (2001) Object permanence in orangutans (*Pongo pygmaeus*), chimpanzees (*Pan troglodytes*), and children (*Homo sapiens*)". J Comp Psychol 115:159–171
Call J (2003) Beyond learning fixed rules and social cues: Abstraction in the social arena. Philosophical Transactions of the Royal Society 358:1189–1196
Call J (2006a) Inferences by exclusion in the great apes: the effect of age and species. Anim Cogn 9:393–403
Call J (2006b) Descartes' two errors: reasoning and reflection from a comparative perspective. In: Hurley S, Nudds M (eds) Rational animals. Oxford University Press, Oxford, pp 219–234
Call J (2007a) Apes know that hidden objects can affect the orientation of other objects. Cognition 105:1–25
Call J (2007b) Past and present challenges in theory of mind research in primates. In: van Hofsten C, Rosander K (eds) Progress in brain research, vol 164. Elsevier, Amsterdam, pp 341–354
Call J (in press). Trapping the minds of apes: causal knowledge and inferential reasoning about objects. In: Lonsdorf E and Ross S (Eds.). Chimpanzee minds. Chicago: The University of Chicago Press
Call J, Carpenter M (2001) Do chimpanzees and children know what they have seen? Anim Cogn 4:207–220
Call J, Tomasello M (2005a) What chimpanzees know about seeing revisited: an explanation of the third kind. In: Eilan N, Hoerl C, McCormack T, Roessler J (eds) Joint attention: communication and other minds. Oxford University Press, Oxford, pp 45–64
Call J, Tomasello M (2005b) Reasoning and thinking in nonhuman primates. In: Holyoak KJ, Morrison RG (eds) Cambridge handbook on thinking and reasoning. Cambridge University Press, Cambridge, pp 607–632
Call J, Tomasello M (2008) Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci 12:187–192
Cunningham CL, Anderson JR, Mootnick AR (2006) Object manipulation to obtain a food reward in hoolock gibbons, *Bunopithecus hoolock*. Anim Behav 71:621–629
Fujita K, Kuroshima H, Asai S (2003) How do tufted capuchin monkeys (*Cebus apella*) understand causality involved in tool use? J Exp Psychol Anim Behav Process 29:233–242
Gillan DJ (1981) Reasoning in the chimpanzee: II. Transitive inference. J Exp Psychol Anim Behav Process 7:150–164
Girndt A, Meier T, Call J (2008) Task constraints mask great apes' proficiency in the trap-table task. J Exp Psychol Anim Behav Process 34:54–62
Gopnik A, Schulz L (2007) Causal learning. Psychology, philosophy, and computation. New York, Oxford University Press, p 358
Hanus D, Call J (2008) Chimpanzees infer the location of a reward based on the effect of its weight. Curr Biol 18(6):R1–2
Harlow HF (1949) The formation of learning sets. Psychol Rev 56:51–65
Hashiya K, Kojima S (2001) Hearing and auditory-visual intermodal recognition in the chimpanzee. In: Matsuzawa T (ed) Primate origins of human cognition and behavior. Springer, Berlin, pp 155–189
Hauser M, Spaulding B (2007) Wild rhesus monkeys generate causal inferences about possible and impossible physical transformations in the absence of experience. Proc Natl Acad Sci USA 103:7181–7185
Hemmi JM, Menzel CR (1995) Foraging strategies of long-tailed macaques, *Macaca fascicularis*: directional extrapolation. Anim Behav 49:457–463
Herrmann E, Call J, Hare B, Hernandez-Lloreda MV, Tomasello M (2007) Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. Science 317:1360–1366
Iriki A, Sakura O (2008) The neuroscience of primate intellectual evolution: natural selection and passive and intentional niche construction. Philosophical Transactions of the Royal Society B 363:2229–2241
Karmiloff-Smith A (1992) Beyond modularity. A developmental perspective on cognitive science. MIT Press, Cambridge, MA

Köhler W (1925) The mentality of apes. Routledge & Kegan Paul, London
Limongelli L, Boysen ST, Visalberghi E (1995) Comprehension of cause-effect relations in a tool-using task by chimpanzees (*Pan troglodytes*). J Comp Psychol 109:18–26
Martin-Ordas G and Call J (in press). Assessing generalization within and between trap tasks in the great apes. Int J Comp Psychol
Martin-Ordas G, Call J, Colmenares F (2008) Tubes, tables and traps: great apes solve two functionally equivalent trap tasks but show no evidence of transfer across tasks. Anim Cogn 11:423–430
Martin-Ordas G, Call J (2009) Assessing generalization within and between trap tasks in the great apes. Inter J of Comp Psychol 22:43–60
Menzel CR (1996) Spontaneous use of matching visual cues during foraging by long-tailed macaques (*Macaca fascicularis*). J Comp Psychol 110:370–376
Mulcahy NJ, Call J (2006) How great apes perform on a modified trap-tube task. Anim Cogn 9:193–199
Natale F, Potì P, Spinozzi G (1988) Development of tool use in a macaque and a gorilla. Primates 29:413–416
Penn D, Povinelli DJ (2007) Causal cognition in human and nonhuman animals: a comparative, critical review. Annu Rev Psychol 58:97–118
Piaget J (Ed.) (1952) The origins of intelligence in children. New York: Norton
Povinelli D (2000) Folk physics for apes: a chimpanzee's theory of how the world works. University Press, Oxford
Premack AJ, Premack D (1983) The mind of an ape. Norton, New York
Premack D, Premack AJ (1994) Levels of causal understanding in chimpanzees and children. Cognition 50:347–362
Seed AM, Call J, Emery NJ, Clayton NS (2009) Chimpanzees solve the trap problem when the confound of tool-use is removed. J Exp Psychol Anim Behav Process 35(1):23–34
Sperber D, Premack D, Premack AJ (1995) Causal cognition. A multidisciplinary debate. Oxford University Press, New York
Spinozzi G, Potì P (1989) Causality I: The support problem. En F. Antinucci (Ed.). Cognitive Structure and Development in Nonhuman Primates (pp. 113–119). Hilldsale (New Jersey): Lawrence Erlbaum Associates
Tebbich S, Bshary R (2004) Cognitive abilities related to tool use in the woodpecker finch, *Cactospiza pallida*. Anim Behav 67:689–697
Tomasello M, Call J (1997) Primate cognition. Oxford University Press, New York
Tomonaga M (1993) Tests for control by exclusion and negative stimulus relations of arbitrary matching to sample in a "symmetry-emergent" chimpanzee. J Exp Anal Behav 59:215–229
Visalberghi E, Limongelli L (1994) Lack of comprehension of cause-effect relations in tool-using capuchin monkeys (*Cebus apella*). J Comp Psychol 108:15–22
Yamamoto J, Asano T (1995) Stimulus equivalence in a chimpanzee (*Pan troglodytes*). Psychol Rec 45:3–21

Part IV
Language, Emotion, Culture

Socializing Cognition

Anne Böckler, Günther Knoblich, and Natalie Sebanz

Abstract The traditional way to study thinking in humans is to investigate cognitive processes in single individuals. The positions laid out in this chapter, by contrast, regard social interaction as the default context within which cognition occurs. The chapter introduces and discusses the theoretical background as well as relevant empirical findings of three approaches that aim at exploring how cognition emerges through and is shaped by social context: 1) Distributed and embodied cognition approaches stress how cognition is inherent in entire socio-technical systems and arises in close interactions between individuals and the environment. 2) Evolutionary and cultural frameworks highlight the role of social interaction in the phylo- and ontogenetic development of higher cognitive functions. 3) Ideomotor approaches postulate close perception-action links and emphasize the contribution of these links for the understanding of other individuals' actions and intentions, implying that perception and action are social by nature. Taken together, the research reviewed in this chapter suggests that respecting the social nature of human cognition will foster a better understanding of individual thinking.

Human cognition is typically studied by focusing on cognitive and brain processes within single minds. For instance, aspects of memory are studied by asking individuals to learn and retrieve lists of words, and processes of action planning and control are studied by asking participants to perform reaction time tasks. In this chapter, we consider what can be gained by "socializing cognition"; taking into account the social context in which cognition occurs, considering the role of evolutionary and cultural forces, and exploring the functionality of perception-action links.

The approaches we are going to describe have in common that they regard social interaction as a sort of default situation in which cognition occurs. The focus is not on the individual's processing of social information, which constitutes a core theme in social cognition research (Smith and Mackie 2001). Rather, the motivation behind these approaches is to explore how cognition emerges through social context and is shaped by it, be it phylogenetically, ontogenetically, or in ongoing interactions.

A. Böckler, G. Knoblich (✉), and N. Sebanz
Centre for Cognition, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, 6500 HB Nijmegen, The Netherlands
e-mail: G.Knoblich@donders.ru.nl

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_16,

Two relatively recent theoretical advances have fuelled cognitive scientists' renewed interest in the mind's connection to its social surroundings. On the one hand, a growing trend towards embodiment emphasizes the role of close interactions between individuals and the environment (Clark 1997). Although there are different notions of embodiment (Wilson 2002), the central claim is that cognition cannot be understood without taking into account the constraints arising when we act in the world in real time. Cognition, according to this approach, needs to be understood in terms of how it contributes to situation-appropriate behavior (Clark 1999). Moreover, not only does the mind serve the body, but also do body and environment serve the mind, for example in the representation and performance of abstract mental tasks (Wilson 2002). In this view, perception, cognition, and action are no longer considered distinct and disconnected, but rather closely interlinked processes that mutually involve each other. This brings about a new point of view: Individual cognition is grounded in the constant interaction of the individual with its environment as well as with other individuals.

Furthermore, evidence in favor of a common system for planning and performing one's own actions and perceiving others' actions (Prinz 1997) has given new impetus to social views on cognition. Activation in the same brain areas when we plan and perform an action as well as when we observe another's action provides evidence for a direct, nonverbal link between people that may support action understanding (Rizolatti and Craighero 2004). The discovery of such perception-action links provides a powerful basis for the investigation of how we understand and anticipate other individuals' actions. Although there is a danger in overestimating what can be achieved through this basic matching principle, it is clear that functionally equivalent representations for self and other provide a crucial platform for integrating the actions of self and other during social interaction. Furthermore, close perception-action links in macaques, as supported by the discovery of mirror neurons that fire during performance and observation of object-directed actions, raise new questions about the evolutionary roots of our cognitive system.

In the following, we will discuss three kinds of approaches that argue for socializing cognition either by distributing cognitive processes across individuals and the environment, by focusing on evolution and culture as shaping forces, or by exploring the functionality of common representations in perception and action.

1 Distributing Cognition Across People and the World

1.1 Distributed Representations

While cognitive and social psychologies tend to focus on individual minds, distributed cognition approaches regard the group as the only meaningful unit of analysis. Cognitive processes are no longer regarded as being bound to the mind and brain

of a single self, but rather as being inherent in the entire socio-technical system (Beer 1995). Cognitive activities are understood primarily as interactions between an agent and physical systems or the agent and other agents (Greeno and Moore 1993). An illustrative example is provided by Hutchins (1995). Imagine a crew in the cockpit of a commercial airliner while approaching to land. In order to successfully accomplish this delicate mission, the crew needs to continuously and precisely compute and remember a large set of correspondences between airspeed and wing configurations. In this context, several technical support systems, the pilots' knowledge, prior experience and memory, but also verbalizations between members of the crew as well as the crew's interactions with technology are, according to Hutchins, part of the effective completion of this task. Therefore, all of these processes must be included in the analysis of cognition in the cockpit. Information is represented and manipulated not only in the pilots' minds (internally) but also by means of technical systems (externally). As cognitive activity occurs between these internal and external representations, the primary unit of analysis is to be the socio-technical system rather than the individual mind.

While this approach has been taken up in applied psychology studies (Rogers and Ellis 1994) and anthropological studies, it has met with little or no response in traditional cognitive psychology and social cognition research. One criticism is that it leads to rather descriptive explanations of phenomena that are specific to the particular situation under investigation, such as interactions in the cockpit. More importantly, once one assumes that cognition is somewhere between people and the world, it becomes unclear how explanations at the level of individual minds fit it. However, embodied approaches to cognition would agree with the general claim that cognition is embedded and cannot be understood without taking into account the interaction between individuals and the environment.

1.2 Coupled Systems

While the distributed cognition approach argues for a new unit of analysis that comprises more than single minds, it does not argue for rejecting the notion of mental representation. Rather, the idea is that mental representations are not necessarily bound to individual minds. In contrast, approaches towards understanding social interaction from an ecological psychology perspective reject the notion of mental representation all together, and argue that principles of dynamical systems should be evoked to explain coupling phenomena between individuals and the environment. Individuals are considered to be perceiving and moving entities whose actions and perceptions dynamically change and reciprocally specify each other (Shaw 2001). The behavior of interacting individuals becomes coupled (entrained) because their perception of the other's actions and their own actions are linked in a way that they mutually affect each other. An example is people's tendency to clap in synchrony (Neda et al. 2000).

The assumption of direct links between perception and action also underlies the key concept of "affordance." Affordances are action opportunities in the environment, and what is perceived as an affordance is thought to depend on one's action abilities. For instance, an object may afford lifting it with one or two hands, depending on one's own hand span (Richardson et al. 2007b). Likewise, when people act in a social context, their arm spans may determine whether an object affords lifting alone or together with others. According to this framework, entrainment and affordance occur on the basis of biological and physical principles and there is no need to involve high-level cognitive representations (Marsh et al. 2006). Hence, the notion of internal representations such as intentions is rejected.

Recent research on entrainment during social interaction has mostly focused on the unintentional coupling of limb movements (Schmidt and Turvey 1994), eye gaze (Richardson and Dale 2005), and body sway (Shockley et al. 2003). For example, when single participants are asked to move two limbs and when two participants are asked to move one limb each as fast as possible, individual and joint movements have been shown to follow the same dynamic principles (Kugler and Turvey 1987; Schmidt et al. 1998). Richardson et al. (2005) showed that participants holding a pendulum in one hand while performing an unrelated verbal task together started swinging their pendulums in synchrony. Moreover, when pairs of participants were sitting side by side in rocking chairs, they unintentionally synchronized rocking speed, even when the eigenfrequencies of the rocking chairs were completely different (Goodman et al. 2005).

There is also recent evidence that synchronization fosters discourse comprehension: The closer the eye movements of a speaker and a listener matched, the better the listener could understand the speaker's information, both when the speaker had previously been videotaped and during online face-to-face conversations (Richardson and Dale 2005; Richardson et al. 2007a). Furthermore, links between synchronization and liking have been reported (Marsh et al., in press). The more people unintentionally synchronize their movements, the more they seem to experience feelings of connectedness.

Generally, the growing interest in applying ecological principles to social behavior provides a basis for new views on how social life shapes cognition. What is attractive about this approach is the strong focus on perception and action that shifts the focus from higher-level cognition to the question of how complex behavior can arise from lower level sensorimotor processes. However, the whole-hearted rejection of mental representations seems rather problematic. As we have argued elsewhere (Sebanz and Knoblich, in press), this considerably limits the range of phenomena to be addressed, and does not seem well suited to explain how individuals flexibly perform tasks together, taking on different roles, and taking each other into account even in the absence of any opportunities for direct exchange.

Taken together, approaches described in this section focus on how cognition is inherent in entire socio-technical systems and how cognition emerges in tight interactions between individuals and the environment. Both the distributed cognition framework and the ecological psychology framework have the potential to shed

new light on our understanding of cognition, reminding us that there is more to cognition than processing stimuli and preparing responses, and that the exploration of human cognition might not be achieved in entirety when social and environmental factors are excluded. Thinking and acting jointly are no longer regarded as the joint product of individually thinking and acting selves. Rather, these approaches take into account the "gestalt quality" of social interaction: the fact that qualitatively different behavior may emerge through social interaction. Furthermore, these approaches stress that cognition in social contexts is not just a special case of regular, individual cognition, but insist that social interaction be considered a "default mode" of action that shapes cognition through and through.

2 Emerging Cognition: Social Brains and Social Norms

The distributed cognition approach and the ecological approach postulate general principles of interaction that can, at least in principle, be applied to all living beings. Ecological theories, in particular, claim that the same principles hold across different species, so that the laws governing flight formation in birds and the laws governing entrainment of movements in humans should be the same. This provides a principled way of addressing "group cognition" across species (Couzin 2008). However, it does not address the relations between the complexity of an individual's social life and his or her cognitive abilities, and differences between species that have evolved over time. How can we account for symbolic communication, the use of Arabic numeral systems, or cultural conventions in human societies? These questions are central to approaches that emphasize the role of social interaction in human development and in the evolution of culture.

2.1 Evolutionary Approaches

Anthropologists and biologists have been busy speculating as to why the neocortex in humans is much larger as compared to our predecessors and as compared to any other species. Humphrey (1976) argued that solving every-day-life problems like finding food, hunting, or evading predators does not necessitate the intellectual capabilities found in some animals and especially humans. Rather, he argued that these cognitive abilities may reflect the need to predict and manipulate conspecifics' behavior when living in social groups. This is in line with the so called "social brain hypothesis" (Dunbar 1998), suggesting that the dramatically increased neo-cortex in the primate and, especially, in the human brain can be traced back to the enlargement of social group size. As groups became larger and more complex (Barton and Dunbar 1997; Sawaguchi and Kudo 1990), the need to form coalitions and maintain relationships increased computational demands

on the individual. Indeed, evidence suggests that the size of the social group is correlated with neo-cortex size across a variety of species (for a review see Dunbar 1998). Abilities like using language, learning from others, and understanding others' mental states ("theory of mind," see Frith and Frith 1999; Frith 2001) are believed to be some of the major effects of increased neo-cortex size, driven by increasing group size (Dunbar 1998; Seyfarth et al. 2005). So far, the focus on evolutionary forces has mainly provided new perspectives on abilities related to social interaction, such as language and theory of mind. However, in principle, it could also provide a new look at cognitive processes that are not typically considered social in nature, such as executive functioning (Roepstorff and Frith 2004) or working memory.

2.2 Cultural Approaches

The same holds for approaches that focus not only on comparative, but also on developmental and cultural aspects of cognition (for a review see Tomasello et al. 2005; Tomasello 1999). Many thoughts in this field go back to Vygotsky's work, emphasizing the role of culture and social interactions in the evolution and the development, respectively, of higher mental functions. According to Vygotsky, a major difference between humans and other primate species lies in the human capability of internalization, a process through which the infant gains knowledge about the signs and symbols of its culture via social interactions with significant adults (Vygotsky 1978). Internalization allows humans to take advantage of others' knowledge and skills, to apply and expand those and to finally shape culture themselves (Vygotsky and Luria 1993).

Comparing a broad range of abilities, including joint attention, imitation, joint action, and mental state attribution in primates, great apes, and humans, Tomasello and colleagues concluded that humans have a unique motivation to share intentions (Tomasello et al. 2005). Although some great apes appear to understand the basics of intentional action (e.g., Hare et al. 2000), they rarely participate in activities that involve imitating another individual, sharing intentions and emotions, or collaborative engagement (e.g., Tomasello and Call 1997; Tomasello 2000). Human infants, by contrast, develop not only the ability to understand others as animate intentional agents during their first 14 month of life, but also develop a species-unique joy and motivation to share psychological states and activities with conspecifics. 1-year old children actively engage in taking turns and role play (Carpenter et al. 2005), as well as in joint attention and pointing, not just in order to achieve certain goals, but for the pure pleasure of it (Liszowski et al. 2004). According to this approach, the unique human motivation to share mental states, to help, learn from, and cooperate with others created the basis for cultural evolution.

Taken together, the theoretical accounts addressed in this section emphasize the role of social interaction in the phylo- and ontogenetic emergence of higher cognitive functions. According to the distributed cognition and ecological

approaches described in the previous section, intelligent behavior emerges in individuals' interactions with each other and the environment. In contrast, comparative approaches allocate cognitive abilities to individual minds that reflect their own and their ancestors' history of social interaction. In the following, we will focus on a third type of theory that centers on the notion of common representations for self and other.

3 Aligning Cognition: Joint Control of Perception–Action Links

During the last decade, researchers in cognitive science and cognitive neuroscience have started to address the social nature of perception–action links, exploring how common mental representations for observed actions, and actions that one can perform oneself could support action understanding, mimicry, imitation, and joint action. This complements the approaches described so far in several ways. Firstly, the interest in perception–action links has led to a focus on basic forms of social interaction occurring in real time, such as people mimicking each other, or coordinating their actions to bring about a change in the environment. This complements the focus on language and theory of mind that generally characterizes cultural and evolutionary approaches. Secondly, the notion of shared action representations for self and others provides a powerful mechanism for a nonverbal form of common ground, allowing individuals to apply their own action knowledge to make sense of others' behavior. This provides a basis for asking how lower-level sensorimotor processes interact with higher-level, intentional planning structures. In contrast to an ecological psychology approach, where the concept of mental representation is rejected in favor of direct, unmediated perception–action links, the key question here is how shared representations emerge, what they entail, and how they are used in the service of intentional action.

3.1 Ideomotor Theories and the Mirror System

According to so-called ideomotor theories, perception and action involve the same mental representations (James 1890; Prinz 1997), because one's own and others' actions are coded in terms of their outcomes or effects. Hence, individuals perceive others' actions in the light of their own action repertoire. Perceiving someone performing an action, such as hearing the sound of hands clapping, automatically activates the action representations the observer would use to perform the observed action himself or herself. This perception–action interface allows us not only to imitate other people's behavior (Meltzoff and Moore 1997) and to predict the outcome of others' actions (Knoblich and Flach 2001; Wilson and Knoblich 2005), but it may also foster the understanding of others' goals and help to infer their intentions (Bekkering et al. 2000; Hamilton and Grafton 2006).

The discovery of "mirror neurons" added a neural substrate to the functional principles postulated by ideomotor theories. These neurons were first localized in the ventral premotor and inferior parietal cortex of macaque monkeys. The core finding is that mirror neurons fire not only when the monkey performs an object-related action (e.g., grasping a peanut), but also when the monkey observes another individual performing the same action (Gallese et al. 1996; Fogassi et al. 2005; for an overview, see Rizzolatti and Craighero 2004).

Electrophysiological and brain imaging studies suggest that there is also a mirror system in humans that serves the same purpose of mapping observed actions to the observer's action repertoire. For instance, the mu-rhythm of the electroencephalogram disappears when one is acting, but also when actions are merely observed (for a review see Rizzolatti et al. 2002). Because the mu-rhythm is known to be suppressed during motor activities, this finding suggests that motor areas participate in action observation. Further evidence comes from brain imaging studies demonstrating that action observation increases metabolic activity in motor areas such as the inferior frontal gyrus, the premotor cortex, and parts of the inferior parietal cortex (Grèzes et al. 2003; Buccino et al. 2004; Rizzolatti and Craighero 2004).

Unlike the primate mirror neuron system, the human mirror system does not only respond to object-directed actions, but also to intransitive movements such as running or dancing (Grèzes et al. 2003). This is especially true when the observed action is well known to the observer such as when a ballet dancer observes another ballet dancer's movements (Calvo-Merino et al. 2005). In the following, we will discuss how perception–action links may contribute to different forms of social interaction.

3.1.1 Mimicry

A wide range of findings on behavioral mimicry suggests that perceiving others' actions automatically induces an unintended tendency in the observer to carry out the same action (Chartrand and Bargh 1999). We unintentionally adopt the bearing (La France, 1979, 1982; Bernieri and Rosenthal 1991), mannerisms (Chartrand and Bargh 1999), and facial expression (Bavelas et al. 1986) of people we are interacting with. In conversations, we tend to adjust the speed and extension of our gestures to each other. For example, we wiggle our feet and fold our arms when our conversation partners do and we adopt their words, clauses, and grammatical structure when we reply to something they said (Bock 1989; Levelt and Kelter 1982; Garrod and Pickering 2004). We also mimic the other's accent and the tone of voice (Giles and Powesland 1975; Neumann and Strack 2000) and we laugh more when we see others laughing (Provine 1992). Although mimicry is omnipresent in social interactions, it is hardly ever explicitly noticed. Rather, people get confused when they don't get mimicked during social interactions (Hatfield et al. 1994).

It is unclear whether all of these mimicry phenomena can be explained through perception–action links. However, at least the mimicking of bodily movements likely

reflects a tendency to perform observed actions due to common representations for self and other (Sebanz and Shiffrar 2006). It is thought that mimicry serves to establish relationships between people; For instance, Chartrand et al. (2005) hypothesized that unconscious mimicry acts as the "social glue that binds and bonds us humans together" (pp 357). In support of this hypothesis they demonstrated that mimicry and mirroring of gestures and postures enhances the perceived fluency of interactions (Chartrand and Bargh 1999). Lakin and Chartrand (2003) found that people who extensively mimic their interaction partner tend to be liked more than those who don't. Interestingly, participants mimicked interaction partners more when they had been rejected beforehand. These findings suggest that unconscious mimicry serves to affiliate with others, to enhance rapport and liking, and to increase the smoothness of interactions.

3.1.2 Imitation

It is likely that the mirror system also plays an important role in the imitation of others' goal-directed actions, which provides a crucial mechanism of cultural learning. Although having a mirror system is not sufficient for the ability to imitate (Knoblich and Sebanz 2008; Pacherie and Dokic 2006), it provides a platform for incorporating others' actions into one's own action system.

One important line of imitation research addresses human infants' imitation abilities. Meltzoff (1990) has demonstrated that 1-year-olds are already able to imitate and also recognize and emotionally value when they are being imitated by others Also, children have the capability to infer from the visible surface of behavior the underlying goals of observed actors. For instance, Meltzoff (1995) made 18-month-old infants watch adults "accidentally" failing to manipulate a target and could show that the infants carried out the intended action, not the unsuccessful attempt. Moreover, children often imitate action goals but do not necessarily use the same means as the model to achieve these goals (Bekkering et al. 2000).

As in mimicry, close perception–action links may provide a matching system for mapping observed actions onto the observer's action repertoire (see Meltzoff and Moore 1997). When adult participants are asked to carry out a grasping movement, they are in fact faster when observing another person performing a corresponding movement (e.g., grasping) as compared to a noncorresponding movement (e.g., spreading) (Stürmer et al. 2000; Brass et al. 2001). However, two additional components seem to be necessary in order to successfully achieve intentional imitation: First, we need to distinguish between self and other (Decety et al. 2002; Decety and Chaminade 2005) in order to keep the intentions of other people and our own intentions apart. Second, we must be able to divide action sequences into meaningful units and remember the right order. As Chaminade et al. (2002) put it: "Imitation is a creative reconstruction of observed action" (p. 327), rather than "monkey see, monkey do".

3.2 *Joint Action*

Much of our daily cognitive activity is embedded in cooperative contexts, where we act together with other people to bring about changes in the environment (Sebanz et al. 2006 a), be it moving furniture together, navigating through traffic, or playing a piano duet. Such joint actions often require that we carry out complementary instead of identical actions (Sebanz et al. 2003; Newman-Norlund et al. 2007). When somebody hands us a glass of wine we need to get hold of it not by imitating the other's hand posture, but by respectively grasping the part of the glass the other person is not touching. This involves forming a representation of one's own and the other's actions and tasks, as well as coordinating actions in time and space. In the following, we will review recent empirical findings to discuss the role of perception–action links for shared representations and coordination in more detail.

3.2.1 Co-representation

Acting together often means carrying out actions that complement each other. We need to complete our part of a task while we see the co-actor completing his or hers. Questions deriving from this fact are how the actions of a co-acting person are mentally represented and how they influence our own actions. In order to examine the representation and influence of another person's complementary action we developed a paradigm – based on the classical Simon task – that allowed us to compare how participants perform one and the same task alone and together (Sebanz et al. 2003). In a Simon task (Simon 1990), participants are asked to respond to nonspatial features of a stimulus (e.g., color) with a spatially arranged response (e.g., left or right button press). For example, they may be asked to respond to red stimuli with a left key press, and to green stimuli with a right key press. An additional spatial feature of the stimulus (e.g., left or right position on the screen) can be compatible or incompatible regarding the response side required by the relevant stimulus feature. Responses are usually faster when relevant and irrelevant stimulus features are compatible as compared to when they are not. For example, red stimuli are easier to respond to with a left button press when they appear on the left than when they appear on the right side of a screen.

According to the dimensional overlap model (Kornblum et al. 1990) this effect is due to the response conflict that arises as the irrelevant spatial stimulus dimension automatically activates the spatially corresponding response. In our study, we used this effect to investigate whether people performing the task together represent each other's actions. The two-choice RT task was either distributed between two people, each responding to one of the two colors (joint condition), or performed by a single person who responded to just one of the colors (individual condition). The same effect reported in the classical setting was found for the joint, but not for the individual condition. In other words, the two hands of two participants acting together produced a similar pattern as the two hands of a single participant performing the whole task.

In contrast, there was no effect when participants performed their part of the task alone (with one hand). This implies that, although the other's action is not relevant for one's own and does not occur at the same time, it is represented in a functionally equivalent way as our own actions and thereby influences our performance.

More generally, this finding provides evidence that close links between perception and action play not only a role during action observation, but can also guide our actions when we perform tasks together with others by taking turns. In line with this view, EEG- recordings revealed an increased no-go P300 amplitude in the joint as compared to the individual condition, pointing towards a larger effort to suppress one's action when it was not one's turn and the stimulus indicated the other's response (Sebanz et al. 2006 b; Tsai et al. 2006).

Several further studies indicate that common representations for one's own and others' actions are controlled by higher-level representations about one's own and others' tasks. People have a tendency to form a representation of their partner's task rules (Sebanz et al. 2005), so that stimuli requiring the other's action elicit an action selection conflict. Furthermore, an fMRI study comparing individual and joint performance suggests that acting together leads to enhanced self-reflective processing and performance monitoring, as co-actors try to make sure whose turn it is (Sebanz et al. 2007). Thus, one could argue that people performing tasks together keep self and other apart at a higher, more abstract level of task representation, but use common action representations for self and other during performance (for more details, see Knoblich and Sebanz 2008).

3.2.2 Coordination

Although forming shared task representations constitutes an important aspect of joint action, many interactions also require tight coordination of actions in space and time. Coordinating our actions with others increases the amount of potential action outcomes and allows for the achievement of performances and products we could not have brought about on our own (Clark 1996). Two equilibrists on swings are an obvious example, jumping, catching, and releasing each other at spiraling speed and height. But also joint actions like rowing a canoe or lifting heavy boxes together require exact and fast predictions and adjustments to the what, when, and where of others' actions (Bosga and Meulenbroek 2007).

Time windows for action coordination often lie in the scope of a few hundred milliseconds, raising the question of how real-time predictions are acquired in joint action (see Sebanz and Knoblich 2008). As described in an earlier section, we unconsciously tend to synchronize our movements with those of others when we interact with them, even when this is not required by the task or when we are asked not to (e.g., Richardson et al. 2005). But how do we achieve intentional and flexible co-timing? One possibility is that forward models (Davidson and Wolpert 2003) that allow us to make temporal predictions about the sensory consequences of our own actions can also be employed when making predictions about others' actions (Wilson and Knoblich 2005).

Predictions of this kind should be most accurate when the observed motor system is identical to ours. This seems to be the case: When participants watched videos of themselves and others' small-scale (handwriting) or large-scale body movements (throwing a dart), they were most accurate in predicting the outcomes of their own actions (Knoblich and Flach 2001; Knoblich et al. 2002). The same holds for auditory recognition: Professional piano players could use temporal cues to detect if they have played a piece themselves or if someone else had, even several months after the pieces had been recorded (Repp and Knoblich 2004). Furthermore, Keller et al. (2007) could demonstrate that pianists duet better when playing along with an earlier recording of themselves as compared to playing along with a recording of another pianist. This can be explained by assuming that temporal predictions are most accurate for one's own performance, thus facilitating synchronization.

Given that two people are never the same, can they ever become as coordinated as a single person? By means of a tracking task that was performed either alone (one person responsible for two response keys) or in dyads (each participant responsible for one of the keys), Knoblich and Jordan (2003) assessed if and under what circumstances two people can coordinate actions as successfully as a single person. With practice, dyads could reach the performance level of single individuals, but only when participants received feedback about each other's actions. This result suggests that one needs to perceive the effects of one's own and others' actions in order to integrate them into predictions about the joint outcome.

Research presented in this section suggests that perception and action are of a "social nature". There is evidence that we understand and predict others' actions by relying on our own motor system. Furthermore, shared action representations are used in the service of joint action. While we need to keep others' intentions (De Lange et al. 2008) and tasks (Sebanz et al. 2007) apart from our own, shared representations provide a platform for predicting and integrating actions of self and other. Like the ecological approach described above, this approach assumes that complex forms of social interaction are grounded in perception–action links. However, a crucial difference is that mental representations are not denied. Rather, shared representations provide a "currency" for social exchange.

4 Conclusion and Outlook

We have described several different approaches to cognition and social interaction that argue for "socializing cognition". The distributed cognition approach takes a radical stance in that it assumes that cognitive processes occur between individuals and the environment. The ecological approach follows this assumption, but replaces information processing concepts with the concepts of dynamical systems theory. Both approaches are strongly embodied and remind us not to neglect tight interactions between individuals and the environment. One could argue, however, that by

rejecting the notion of individual cognition and mental representation, cognition is "socialized" to the extent that it dissolves in interaction. This makes it difficult to approach phenomena that seem to rely on people's thoughts about each other.

Evolutionary and cultural approaches are more obviously, but perhaps less radically, social. Claims about the role of evolutionary and cultural forces focus on higher-level cognitive abilities deemed critical for social interaction, such as theory of mind. Thus, cognition is "socialized" by emphasizing the power of social interaction. However, it remains open to what extent social forces shape lower-level cognitive functions, and how higher-level cognition may be grounded in lower-level processes.

Finally, ideomotor theories and the notions of common coding, shared representations, and "mirroring" can be regarded as providing new pieces to a puzzle rather than delivering a unified theory for socializing cognition. The crucial insight gained from this approach is that perception and action are not just servants to mighty minds that process information to compute outputs. Rather, tight links between perception and action enable a range of abilities that were previously poorly understood, such as mimicry, imitation, and joint action. One limitation of this approach is that it is largely based on evidence from studies on action perception. We still know rather little about the role of perception–action links in joint action, and more generally, on how higher-level planning processes mediate perception–action links. Future research on these topics will likely contribute to our understanding of cognition in joint action.

Taking a joint perspective may change not only our understanding of action planning and control, but can also provide new perspectives on other cognitive processes, such as memory, problem solving (Wooldridge and Jennings 1999), and categorization (Markman and Makin 1998). For instance, in the memory domain, joint cognition has been addressed by studying "transactive memory." This refers to the encoding, storing and retrieving of information shared by two or more people (Wegner 1986). Research on transactive memory demonstrates that people use each other's minds as resources, particularly when they know each other's minds well. For instance, elderly long-term partners were shown to outperform randomly selected elderly pairs of people in a wide range of memory tasks (Johansson et al. 2000). A study by Wegner et al. (1991) showed that transactive memory supports performance when pairs can choose their own recall strategy, but is impaired when a dictated strategy requires reorganization and reassignment of their transactive memory system.

Synergies between minds working together are also known with respect to problem solving and joint thinking. Perhaps the most famous example to illustrate successful joint thinking is research. (e.g., Tversky and Kahneman 1983, 1982; Kahneman and Tversky 1996). As Nobel laureate Kahneman pointed out, statistical evidence confirmed that their joint work was of higher quality and more influential than the work they accomplished individually (Laibson and Zeckhauser 1998). He explained "Amos and I shared the wonder of together owning a goose that could lay golden eggs – a joint mind that was better than our separate minds." (Kahneman 2002). Future research will hopefully address the basis of such synergy effects in joint cognition.

Finally, taking the idea of "socializing cognition" seriously may change our views on cognitive processes that are considered to be independent of or even impenetrable by social context. As cognitive (neuro)science experiments are rarely carried out in entirely nonsocial situations, some cognitive phenomena established by studying single individuals might well turn out to be mediated by social factors once the social context is systematically varied. A striking example is a recent study of perseveration errors (A-not-B error) in 10-month-old infants (Topál et al. 2008).

When an object is hidden in one of two locations, infants tend to search the location where they repeatedly retrieved the object before, even when they observed the object being hidden in the other location in the actual trial (Piaget 1954). The cognitive process underlying this error was mostly suggested to be the infant's inability to inhibit previously rewarded responses. By contrast, Topal and colleagues demonstrated that communicative cues that are provided by the experimenter cause an interpretation bias that misleads infants in the classical version of the task. When infants did not see the experimenter, their performance significantly improved. So, what was thought to be a failure of individual minds seems to at least partly reflect a social process that normally helps infants to learn from communication. This supports our conclusion that considering the social nature of the human mind and brain can help us to better understand the nature of individual cognition itself.

References

Barton RA, Dunbar RLM (1997) Evolution of the social brain. In: Whiten A, Byrne R (eds) Machiavellian intelligence, vol 2. Cambridge University Press, Cambridge

Bavelas JB, Black A, Lemery CR, Mullett J (1986) "I show how you feel": motor mimicry as a communicative act. J Pers Soc Psychol 50:322–329

Beer RD (1995) A dynamical systems perspective on agent–environment interaction. Artif Intell 72(1–2):173–215

Bekkering H, Wohlschläger A, Gattis MI (2000) Imitation of gestures in children is goal-directed. Q J Exp Psychol 53:153–164

Bernieri FJ, Rosenthal R (1991) Interpersonal coordination: behavior matching and interactional synchrony. In: Feldman RS, Rimé B (eds) Fundamentals of nonverbal behavior. Cambridge University Press, Cambridge, England, pp 401–432

Bock JK (1989) Closed-class immanence in sentence production. Cognition 31:355–387

Bosga J, Meulenbroek RGJ (2007) Joint-action coordination of redundant force contributions in a virtual lifting task. Motor Control 11:234–257

Brass M, Bekkering H, Prinz W (2001) Movement observation affects movement execution in a simple response task. Acta Psychol 106(1–2):3–22

Buccino G, Binkofski F, Riggio L (2004) The mirror neuron system and action recognition. Brain Lang 89(2):370–376

Calvo-Merino B, Glaser DE, Grezes J, Passingham RE, Haggard P (2005) Action observation and acquired motor skills: an fMRI study with expert dancers. Cereb Cortex 15(8):1243–1249

Carpenter M, Tomasello M, Striano T (2005) Role reversal imitation and language in typically developing infants and children with autism. Infancy 8(3):253–278

Chaminade T, Meltzoff AN, Decety J (2002) Does the end justify the means? A PET exploration of the mechanisms involved in human imitation. Neuroimage 15(2):318–328

Chartrand TL, Bargh JA (1999) The chameleon effect: the perception-behavior link and social interaction. J Pers Soc Psychol 71:464–478
Chartrand TL, Maddux WW, Lakin JL (2005) Beyond the perception-behavior link: the ubiquitous utility and motivational moderators of nonconscious mimicry. In: Hassin RR, Uleman JS, Bargh JA (eds) The new unconscious. Oxford University Press, New York, pp 334–361
Clark A (1997) Being there: putting brain, body, and world together again. MIT Press, Cambridge, MA
Clark A (1999) An embodied cognitive science? Trends Cogn Sci 3(9):345–351
Clark HH (1996) Using language. Cambridge University Press, Cambridge, UK
Couzin ID (2008) Collective cognition in animal groups. Trends Cogn Sci 13:36–43
Davidson PR, Wolpert DM (2003) Motor learning and prediction in a variable environment. Curr Opin Neurobiol 13(2):232–237
Decety J, Chaminade T, Grèzes J, Meltzoff AN (2002) A PET exploration of the neural mechanisms involved in reciprocal imitation. Neuroimage 15(1):265–272
Decety J, Chaminade T (2005) The neurophysiology of imitation and intersubjectivity. In: Hurley S, Chater N (eds) Perspectives on imitation: from cognitive neuroscience to social science, vol 1. MIT Press, Cambridge, pp 119–140
Dunbar RIM (1998) The social brain hypothesis. Evol Anthropol 6:178–190
Fogassi L, Ferrari PF, Gesierich B, Rozzi S, Chersi F, Rizzolatti G (2005) Parietal lobe: from action organization to intention understanding. Science 308(5722):662–667
Frith CD, Frith U (1999) Interacting minds – a biological basis. Science 286(5445):1692–1695
Frith U (2001) Mind blindness and the brain in autism. Neuron 32(6):969–979
Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain 119(2):593–609
Garrod S, Pickering MJ (2004) Why is conversation so easy? Trends Cogn Sci 8(1):8–11
Giles H, Powesland PF (1975) Speech style and social evaluation. Academic Press, London
Goodman JRL, Isenhower RW, Marsh KL, Schmidt RC, Richardson MJ (2005) The interpersonal phase entrainment of rocking chair movements. In: Heft H, Marsh KL (eds) Studies in perception and action VIII: Thirteenth International conference on perception and action. Lawrence Erlbaum Associates, Inc, Mahwah, NJ, pp 49–53
Greeno JG, Moore JL (1993) Situativity and symbols: response to vera and simon. Cogn Sci 17(1):49–59
Grèzes J, Armony JL, Rowe J, Passingham RE (2003) Activations related to "mirror" and "canonical" neurons in the human brain: an fMRI study. Neuroimage 18(4):928–937
Hamilton AF, Grafton ST (2006) Goal representation in human anterior intraparietal sulcus. J Neurosci 26(4):1133–1137
Hare B, Call J, Agnetta B, Tomasello M (2000) Chimpanzees know what conspecifics do and do not see. Anim Behav 59(4):771–785
Hatfield E, Cacioppo JT, Rapson RL (1994) Emotional contagion. Cambridge University Press, Cambridge
Humphrey NK (1976) The social function of intellect. In: Bateson PPG, Hinde RA (eds) Growing points in ethology. Cambridge University Press, Cambridge, UK, pp 303–317
Hutchins E (1995) How a cockpit remembers its speeds. Cognit Sci: A Multidisciplinary Journal 19(3):265–288
James W (1890) The principles of psychology. Holt, New York, NY
Johansson O, Andersson J, Roenneberg J (2000) Do elderly couples have a better prospective memory than other elderly people when they collaborate? Appl Cogn Psychol 14:121–133
Kahneman D, Tversky A (1996) On the reality of cognitive illusions: a reply to Gigerenzer's critique. Psychol Rev 103:582–591
Kahneman D (2002) Autobiography. In: Frängsmyr T (Ed.) Les Prix Nobel. The Nobel Prizes 2002, Stockholm, 2003. (http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahneman-autobio.html)
Keller P, Knoblich G, Repp BH (2007) Pianists duet better when they play with themselves. Conscious Cogn 16:102–111

Knoblich G, Flach R (2001) Predicting the effects of actions: interactions of perception and action. Psychol Sci 12(6):467–472
Knoblich G, Seigerschmidt E, Flach R, Prinz W (2002) Authorship effects in the prediction of handwriting strokes: evidence for action simulation during action perception. Q J Exp Psychol A55:1027–1046
Knoblich G, Jordan JS (2003) Action coordination in groups and individuals: learning anticipatory control. J Exp Psychol Learn Mem Cogn 29(5):1006–1016
Knoblich G, Sebanz N (2008) Evolving intentions for social interaction: from entrainment to joint action. Philos T Roy Soc B: Biol Sci 363(1499):2021–2031
Kornblum S, Hasbroucq T, Osman A (1990) Dimensional overlap: cognitive basis for stimulus–response compatibility: a model and taxonomy. Psychol Rev 97:253–270
Kugler PN, Turvey MT (1987) Information, natural law and the selfassembly of rhythmic movement. Erlbaum, Hillsdale, NJ
La France M (1979) Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. Soc Psychol Quart 42:66–70
La France M (1982) Posture mirroring and rapport. In: Davis M (ed) Interaction rhythms: periodicity in communicative behavior. Human Sciences Press, New York, pp 279–298
Laibson D, Zeckhauser R (1998) Amos Tversky and the ascent of behavioral economics. J Risk Uncertainty 16:7–47
Lakin JL, Chartrand TL (2003) Using nonconscious behavioral mimicry to create affiliation and rapport. Psychol Sci 14(4):334–339
De Lange FP, Spronk M, Willems RM, Toni I, Bekkering H (2008) Complementary systems for understanding action intentions. Curr Biol 18:454–457
Levelt WJM, Kelter S (1982) Surface form and memory in question answering. Cogn Psychol 14:78–106
Liszowski U, Carpenter M, Henning A, Striano T, Tomasello M (2004) Twelve-month-olds point to share attention and interest. Dev Sci 7:297–307
Markman AB, Makin VS (1998) Referential communication and category acquisition. J Exp Psychol Gen 127:331–354
Marsh KL, Richardson MJ, Baron RM, Schmidt RC (2006) Contrasting approaches to perceiving and acting with others. Ecol Psychol 18(1):1–38
Marsh K, Richardson MJ, Schmidt RC (2009) Social connection through joint action and social synchrony. Top Cogn Sci 1(2):320–339
Meltzoff AN (1990) Foundations for developing a concept of self: the role of imitation in relating self to other and the value of social mirroring, social modeling, and self practice in infancy. In: Cicchetti D, Beeghly M (eds) The self in transition: infancy to childhood. University of Chicago Press, Chicago, pp 139–164
Meltzoff AN (1995) Understanding the intentions of others: re-enactment of intended acts by 18-months-old children. Dev Psychol 31:838–850
Meltzoff AN, Moore MK (1997) Explaining facial imitation: a theoretical model. Early Dev Parenting 6:179–192
Neda Z, Ravasz E, Brechet Y, Vicsek T, Barabasi A-L (2000) Self-organizing processes: the sound of many hands clapping. Nature 403:849–850
Neumann R, Strack F (2000) "Mood contagion": the automatic transfer of mood between persons. J Pers Soc Psychol 79(2):211–223
Newman-Norlund RD, van Schie HT, van Zuijlen AMJ, Bekkering H (2007) The mirror neuron system is more active during complementary compared with imitative action. Nat Neurosci 10:817–818
Pacherie E, Dokic J (2006) From mirror neurons to joint actions. Cogn Sys Res 7:101–112
Piaget J (1954) The construction of reality in the child. Basic Books, New York
Prinz W (1997) Perception and action planning. Eur J Cogn Psychol 9:129–154
Provine RR (1992) Contagious laughter: laughter is a sufficient stimulus for laughs and smiles. Bull Psychon Soc 30:1–4
Repp BH, Knoblich G (2004) Perceiving action identity. Psychol Sci 15(9):604–609

Richardson DC, Dale R (2005) Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cognit Sci: A Multidisciplinary Journal 29(6):1045–1060
Richardson MJ, Marsh KL, Schmidt RC (2005) Effects of visual and verbal interaction on unintentional interpersonal coordination. J Exp Psychol Hum Percept Perform 31(1):62–79
Richardson DC, Dale R, Kirkham NZ (2007a) The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. Psychol Sci 18(5):407–413
Richardson MJ, Marsh KL, Baron R (2007b) Judging and actualizing intrapersonal and interpersonal affordances. J Exp Psychol Hum Percept Perform 33:845–859
Rizzolatti G, Fogassi L, Gallese V (2002) Motor and cognitive functions of the ventral premotor cortex. Curr Opin Neurobiol 12(2):149–154
Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annu Rev Neurosci 27(1):169–192
Rogers Y, Ellis J (1994) Distributed cognition: an alternative framework for analysing and explaining collaborative working. J Inform Technol 9:119–128
Sawaguchi T, Kudo H (1990) Neocortical development and social structure in primates. Primates 31(2):283–289
Schmidt RC, Bienvenu M, Fitzpatrick PA, Amazeen PG (1998) A comparison of intra- and interpersonal interlimb coordination: coordination breakdowns and coupling strength. J Exp Psychol Hum Percept Perform 24:884–900
Schmidt RC, Turvey MT (1994) Phase-entrainment dynamics of visually coupled rhythmic movements. Biol Cybern 70:369–376
Sebanz N, Knoblich G, Prinz W (2003) Representing others' actions: just like one's own? Cognition 88(3):11–21
Sebanz N, Knoblich G, Prinz W (2005) How two share a task: corepresenting stimulus-response mappings. J Exp Psychol Hum Percept Perform 31(6):1234–1246
Sebanz N, Bekkering H, Knoblich G (2006a) Joint action: bodies and minds moving together. Trends Cogn Sci 10(2):70–76
Sebanz N, Knoblich G, Prinz W, Wascher E (2006b) Twin peaks: an erp study of action planning and control in coacting individuals. J Cogn Neurosci 18(5):859–870
Sebanz N, Rebbechi D, Knoblich G, Prinz W, Frith CD (2007) Is it really my turn? An event-related fMRI study of task sharing. Soc Neurosci 2(2):81–95
Sebanz N and Shiffrar M (2006) Bluffing bodies: inferring intentions from actions. EPS Conference Proceedings
Sebanz N, Knoblich G (2008) From mirroring to joint action. In: Wachsmuth I, Lenzen M, Knoblich G (eds) Embodied communication. Oxford University Press, Oxford, pp 129–150
Sebanz N and Knoblich G (in press) Jumping on the ecological bandwagon? Mind the gap! *Eur J Soc Psychol*
Seyfarth RM, Cheney DL, Bergman TJ (2005) Primate social cognition and the origins of language. Trends Cogn Sci 9:264–266
Shaw RE (2001) Processes, acts, and experiences: three stances on the problem of intentionality. Ecol Psychol 13:275–314
Shockley K, Santana MV, Fowler CA (2003) Mutual interpersonal postural constraints are involved in cooperative conversation. J Exp Psychol Hum Percept Perform 29:326–332
Simon JR (1990) The effects of an irrelevant directional cue on human information processing. In: Proctor RW, Reeve TG (eds) Stimulus-response compatibility: an integrated perspective. North-Holland, Amsterdam, pp 31–86
Smith ER, Mackie DM (1999) Social psychology. Psychology Press, Philadelphia
Stürmer B, Aschersleben G, Prinz W (2000) Correspondence effects with manual gestures and postures: a study of imitation. J Exp Psychol Hum Percept Perform 26:1746–1759
Tomasello M, Call J (1997) Primate cognition. Oxford University Press, New York
Tomasello M (1999) The cultural origins of human cognition. Harvard University Press, Cambridge
Tomasello M (2000) Culture and Cognitive development. Curr Dir Psychol Sci 9:37–40
Tomasello M, Carpenter M, Call J, Behne T, Moll H (2005) Understanding and sharing intentions: the origins of cultural cognition. Behav Brain Sci 28:675–691

Topál J, Gergely G, Miklósi A, Erdöhegyi A, Csibra G (2008) Infants' perseverative search errors are induced by pragmatic misinterpretation. Science 321:1831–1834

Tsai C-C, Kuo W-J, Jing J-T, Hung D, Tzeng O (2006) A common coding framework in self–other interaction: evidence from joint action task. Exp Brain Res 175(2):353–362

Tversky A, Kahneman D (1982) Evidential impact of base rates. In: Kahneman D, Slovic P, Tversky A (eds) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge, UK, pp 153–160

Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol Rev 90:293–315

Vygotsky L (1978) Interaction between learning and development. In: Vygotsky L (ed) Mind in society. Harvard University Press, Cambridge, MA, pp 79–91

Vygotsky LS and Luria AR (1993) Studies on the history of behavior: ape, primitive and child Golod VI and Knox JE (Eds. and Trans.). Hillsdale, NJ: Lawrence Erlbaum Associates. Original published in 1930

Wegner DM (1986) Transactive memory: a contemporary analysis of the group mind. In: Mullen B, Goethals GR (eds) Theories of group behavior. Springer-Verlag, New York, pp 185–208

Wegner DM, Raymond P, Erber R (1991) Transactive memory in close relationships. J Pers Soc Psychol 61:923–929

Wilson M (2002) Six views of embodied cognition. Psychon Bull and Rev 9:625–636

Wilson M, Knoblich G (2005) The case for motor involvement in perceiving conspecifics. Psychol Bull 131:460–473

Wooldridge M, Jennings NR (1999) Cooperative problem solving. J Logic Comput 9(4):563–592

Thinking and Language

A View from Cognitive Semio-Linguistics

Per Aage Brandt

Abstract Cognitive semio-linguistics studies the relations between signs and language, that is, between semiological and linguistic structures, as expressions of, and as causes of, the cognitive activities involved in thinking, here called epistemic activities. This short essay displays a leveled analysis of the relations holding between semio-linguistic and epistemic structures active in the human mind.

1 Semiotic Bridges

Language – spoken, written or signed – is likely to be the main bridge between *communication* and *cognition* in our species. At one end of this bridge, we find a display of temporally or graphically linear flows of signs ("strings") grouped into words and sentences that are shared, whether immediately or through mediating devices, by shifting speakers and hearers, as meaningful discourse – as debate, dialog, monolog, or text. At the other end of the bridge, individual human agents are each in their singular, embodied, isolated minds attending to concrete or abstract personal or communal matters that call for thinking, imagining, feeling, planning, acting – and also call for being linguistically expressed. The result is the community of beings that communicate important parts of their thinking and which we call culture, civilization, and humanity.

However, the communicational end of the bridge is also a world of traces, signals, and images, that is, of many sorts of non-linguistic but still interpretable and meaningful signifiers that we constantly produce and perceive, and with which our linguistic signs compete and combine. We thus live in a universe of signs, and the linguistic flows

P.A. Brandt
College of Arts & Sciences, Case Western Reserve University, 10900 Euclid Avenue, 44106-7068, Cleveland, OH, USA
e-mail: pab18@case.edu; peraage.brandt@case.edu

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_17,

of signs are often submerged by other significant semiotic flows, often also conceptually efficient, and often vitally urgent.

Consciousness famously experiences *itself* as happening *in the outer world* (not "in the head").[1] This basic phenomenological fact lets our own constant flows of external expressive doings (also happening in the outer world) be experienced as directly connected to, and even identified with, the abstract or concrete matters that we think about and attend to. We experience things, concepts, and signs with equally salient force and as given and co-present in the same outer world. Our minds naturally feel that, for example, things, concepts of things, and signs of concepts of things are aspects of the same reality and are real entities. Only an evolutionarily late, historical development of theory (philosophy) has allowed us to distinguish these aspects and understand their relative independence, so that we can ask questions about the relations between signs in general, language, thinking, and the reality that thinking refers to.

There is a particular kind of culturally developed conventional signs that we need to consider: *writing*. We write mathematical, musical, and linguistic texts. Such writing is called *symbolic*; it is in general "digital", that is, performed in finger and hand scale. Additionally, we draw and paint *images*, that is, produce *iconic* representations. And perhaps even more interestingly, we use *diagrams* to express our thinking; these diagrammatic, spontaneously half-symbolic, half-iconic forms of graphic activity apparently cover all domains of possible thinking.[2]

The cognitive feedback that our minds receive from external symbolic, iconic, and diagrammatic representations is massive and decisive; it deeply influences our thinking and shapes our views of reality, perhaps even more profoundly than the live experience of the situations we are in.

Most often, language does not express pure thinking, but instead interprets these representations that already represent thinking. Symbols, icons, and diagrams (as well as traces, symptoms, signals, etc.; we will here call the whole of non-linguistic signs *semiological*) are apparently spontaneous productions of the human mind that constitute a shorter bridge between communication and cognition than the one offered by language. These semiological expressions are apparently, and probably, both more directly connected to the process of thinking and more directly shaped by the *structure* of the epistemic activity, the process of thinking. However, it takes a certain amount of transcription, translation, and paraphrasis in terms of a human verbal language to make sense of these more "authentic" symbolico-iconic signs of our thinking. It takes a linguistic interpretation of our signs to *socially* transmit their meaning; the variations occurring in our linguistic interpretation of our own and

[1] This is the basic point made by Maurice Merleau-Ponty (1960). The essay "Sur la phénoménologie du langage" (1951, in 1960) is particularly important to the present analysis.

[2] In view of the immense conceptual range of diagrammatic representations, the study of the natural "logic" of diagrams is an important task for a cognitive semiotics (see "The Semantics of Diagrams", Brandt (2004)).

each other's non-linguistic expressions may even explain the dynamic and creative character of thoughtful communication – and hence certain aspects of the history of ideas.[3]

Language constitutes a longer, slower, and more complex bridge between communication and cognition; however, since both the shorter (semiological) and the longer (linguistic) bridge offer essential advantages[4], they are both in constant use, and the "dialectics" – in the sense of competition and conflict, but also of coordination and mutual interpretation – between the two semiotic bridges, the linguistic and the semiological, determine the main expressive and creative functioning of the mind.

2 Levels of Language Structure

Let us consider what is actually and currently known about linguistic structure.[5] We will have to distinguish two general directions: one is productive, *efferent*, beginning in the cognitive, epistemic process of perceiving and thinking, and ending in expressive, prosodic, and linearized discourse; the other is receptive, *afferent*, starting in the cognitive apperception of the linear string of discourse and ending in a contribution to the multidimensional process of thinking. To understand (decode) linguistic expression, as a hearer or a reader, and to express (encode) one's own thinking, as a speaker or a writer, are distinct things, and we cannot assume that one process is exactly the inverse of the other, sharing all structural instances and mechanisms implied. However, it seems sound to assume that core structures underlying both processes, the efferent and the afferent, are indeed shared.

In so far as we can in principle consider a given linguistic manifestation as an expression of thinking, and thinking as the content of this linguistic expression, we can in fact envision the field as a long semiotic bridge allowing two-way traffic between a phonetics (in a large sense) and a semantics of thinking (in a large sense).

I will here boldly present an architectural model of this bridge, or processual network. I will isolate five grounding structural stances or levels that we have to distinguish, as a minimum, and that we have to conceive as connected, locally and transversally, in the "logo-phonic" articulation of the semiotic bridge.

---

[3] Yuri Lotman (1990) made a similar observation, suggesting that the mutual translation of irreducibly different semiotic systems is a core principle in meaning production.

[4] Linguistic representation of meaning is closely related to intersubjective contact and affective communication, because it offers nuanced emotional information; semiological representation is more closely and directly related to (pure) thinking, precisely because it does not convey such emotional information.

[5] We will comfortably ignore the strifes of theories of language that has characterized the history of linguistics from its origins in nineteenth century philology to its present agony in the arms of computer science.

3 Five Logophonic Pillars

The one-dimensional phonetic or graphic expression of language, consisting of phonemes in syllables and the latter grouped in syllabic clusters corresponding to phrases composed of words, and groups of phrases in prosodic ensembles, is a field of highly complex phenomena, acquired procedurally in first language development and (partially, at the cost of a "foreign" accent) acquired through conscious training in subsequent language learning.
We will call this level phonetics:

I. *Phonetics*: Linear structure, concatenation of phonemic, syllabic, lexemic and morphemic entities under a prosodic profile. In "1D".[6]

Parsing is the natural process of reading off the string of phonetic manifestations and reorganizing its parts in a grammatical pattern of connected phrases. Linearization is the inverse process of projecting grammatical structure onto a one-dimensional string. The organization of these "parts of discourse" (French: "les parties du discours") in networks of interconnected meaningful phrases is known from school grammars using varying descriptive terminologies that simply rely on the learners' intuitions; we all possess such intuitions, to a certain degree, allowing us to find the "immediate constituents", the finite verb and the main nominal complements in a sentence, and then to interpret some of the morphological and adverbial meanings that help the ensemble make sense. We will call this level grammar:

II. *Grammar*: syntactic node-structure ("tree structure"), with semantically significant nodes, accounting for meaningful constituent assembly; verbal networks, or constituent "trees", will embed nominal networks, and the highest level will present itself as a "tree of trees", in which embedded phrases and clauses end in a matrix *sentence* carrying the main tone of information conveyed by the utterance.[7]

The format is a network of nodes embedding other networks of nodes according to a canonical node semantics that lets complementation add information to phrase heads. This format is necessarily at least two-dimensional (a node is a bifurcation read backwards). The diagrammatic model will spontaneously be schematized through the verticality of hierarchy, hypotaxis, and the horizontality of coordination, parataxis. Morphological meaning then runs in both directions, sideways (for example, between nouns and determiners) and perpendicularly (for example, from verb to adverb or to subordinate clause).[8] It is "2D".

---

[6] 1D: One dimension. The decoding process can be seen as a "funnel" leading from 1D to 3D and 3D+ structures, through the 2D structures of syntax.

[7] Tone of information: mode of enunciation – volitive, interrogative, assertive, affective, or other (ironic, quotative, etc.).

[8] My own more special theory of so-called *stemmatic* grammar is briefly summarized in Brandt 2004 et passim. It builds on the discovery that the *semantics of syntactic nodes* is schematic and canonical: a short list of semantically informed nodes form canonical cascades that allow recursion and thereby establishes our capacity to spontaneously create and immediately grasp even very complex syntactic networks as meaningful. This discovery solves the problem of defining case structure in a finite and manageable way.

The advantage of the dimensional shift is evident: meaning is compressed or decompressed between 2D and 1D representations. It is further decompressed into – or compressed from – the third level we will consider.

Sentences are grammatically meaningful units, and their information will constantly refer to larger situational semantic frames structured by complementary information in a more general format. Most situational meaning portions refer to parts of a composition like the following, which we may call a natural proposition: *Agents Accessing Objects and Modifying them in view of some final Destination* (A access O, achieving O –> O*, with goal D), combined with other Agents etc. The classical example is a restaurant "script" or a selling-and-buying (or teaching-and-learning) frame encompassing multiple interactions of this sort.[9] There simply *must* be a frame-organized semantics behind sentences, since we can paraphrase, rephrase, "window" in and out components of syntactic structures, expand and contract, and we can translate from language to language while changing the source syntactic structure; in these cases, while allowing differences between source structure and target structure, we ideally maintain the "underlying" frame of meaning, the natural proposition that a sentence represents.[10]
We will call this level semantics:

III. *Semantics*: an event is conceptualized in a situational frame structure, and its information is dimensioned in view of accounting for agency, motion, change, and exchange. This level structures semi-equivalences between different syntactic structures, such as active and passive, or verbal and nominal representations of the same event. It integrates lexical entities (words of word classes). Since it is situational, or episodic, it is "3D".[11]

Events and their frames are further, or previously, necessarily understood as meaningful on the background of general knowledge of the domains of experience to which they may belong.[12] We know from conceptual metaphor that semantic domains are "tectonic" underlying regions of experiential meaning that cultures fill with items but which share constitutive boundaries: physical (D1), social (D2), mental (D3), and communicational (D4) experiences are cognized in different conceptual formats. Causality (D1), finality (D2), intentionality (D3), volition (D4), are all

[9]Schank and Abelson 1977; Fillmore and Atkins 1992. Literature on frames and scripts is extensive, although the problem of formatting the frames has not been solved.

[10]Croft (2007) and Chafe (2005) refer to semantics, or meaning, as a whole of experience that grammar partializes, and interpretation retotalizes. I consider this view as cognitively insufficient; utterance meaning rather represents situational meaning, which further represents knowledge-based epistemic meaning, which represents the embodied process of thinking itself. Meaning is a stack of representations, inside and outside of language.

[11]Root words that label categories are of course linked to level III structure, since they are constant components under variation of possible grammatical structures, and they are core components of frames.

[12]Meaning through language is thus both "shallow level" (flat) and still d"eep" and encyclopedic, that is, rooted in long-term memorized knowledge. Hagoort and van Berkum (2007) show that in fact world knowledge is immediately activated in sentence decoding.

versions of *forces* but give rise to separate forms of experience, which then combine in intricate "higher" order concepts in human cognition.[13] Knowledge is organized in our memory under distinct domain headings; terminologies differ from domain to domain, which precisely is what allows metaphorical transfers. Natural "habits" (D1) are not social deontic "rules" (D2), and mental items like "beliefs" (D3) are not assertions or postulations (D4). We will call this level phenomenological:

IV. *Phenomenology*: experiential domains of concepts memorized by speaker or hearer, ideally shared by speakers; offering an encyclopedic referential background of semantic frames and articulated into regions of possible, generic, or factual knowledge (physical, social, mental, communicational, and of higher orders).

Finally, we have to acknowledge the relevance of a level of meaning constituted by the thinking itself. The human subject incorporating the agent of thinking is a *person* and is typically having a so-called *problem*, practical or theoretical. In a given context, a part of the mental "landscape" of the subject is unclear and triggers a quest for clarity. Negations in language ("I don't see…") refer fundamentally to this phenomenon: a *local lack of clarity* within a larger context of better conceptualized or identified states of affairs and known circumstances. The unclear subregion can of course be of any domain, or of several domains, or still undetermined as to its domain (is the unknown cause of an undesired situation physical, social, mental, communicational, or other?). The thinker or addressee of thinking is typically situated in a real context that allows the fixation and circumscription of the unclear subregion to take on special relevance; the subject is, in some sense, "in trouble". We think when we are in trouble. Human minds in fact seem to prefer to stay in some forms of "trouble", in order to stay in the mode of thinking. This mode possesses modal characteristics: what the subject wants or has to do, the subject cannot do; or what the subject wants to do, the subject can but must not do; or what the subject must do, it can but wants not to do. The unrealistic, "happy" subject, by contrast, would be in the following situation: all it must do it also wants to do, and all it wants to do it also can do; but still it may do things it should not, and thus be "in trouble". Modalities and feelings, including emotions, are essential aspects of thinking. These aspects of thinking will influence the overlayering structure and eventually turn up as distinguishable properties of phonetic prosody. We know from narratives and history in general that "being in trouble" is a core condition of intellectual and epistemic activity. We will call the corresponding, last level epistemic.

V. *Epistemic structure*: The speaker's and/or hearer's actual topic for thinking, related to a narrative that circumscribes the situation of thinking and speaking.

All levels are connected, both serially (I↔II↔III↔IV↔V) and transversally, e.g. II↔V for enunciational modes: volitive, interrogative, assertive, exclamative.

[13]A theory of such semantic or experiential domains, also called ontological domains, is given in Brandt 2004. Temporal cognition schemas are different in the basic domains D1–D4, and object categories are essentially different. *Others* are characteristically distinct: everybody (D1), we/they (D2), I myself (D3), you whom I am addressing (D4).

It is difficult to decide where linguistics ends and cognition begins, in this architecture of structures that communicate in complex ways. I↔II↔III form a "dimensional funnel" 1D through 3D. Processes at III↔IV↔V anchor mental representations in referential meanings, and the relevance of the latter in thinking as a search for "clarity" in the sense of substituting conceptual contents for circumscribed voids. To summarize:

I. **Phonetics** (temporal or otherwise linear manifestations of expressions)
II. **Grammar** (networks of functional constituents linked by semantic nodes)
III. **Semantics** (in terms of frames of groups of worded natural propositions)
IV. **Phenomenology** (experience and knowledge organized in domains)
V. **Epistemic** activity (thinking proper, based on problems or "trouble")

Since we think in situations we are in that are also situations of communication – with ourselves or with others, at a variable distance – *the epistemic activity is also a pragmatic activity*: thinking is doing, and doing is a normal context for thinking; the epistemic level is therefore the *pragmatic* level. Instead of a five-level cake model, we may thus consider a spiral representation as the following Fig. 1.

Speaking leads to phonation, undeniably: V → I. But the Thinking part of the pragmatic activity has to imply a process working backwards from V to I: Thinking (V) combines memorized knowledge pertaining to different domains (IV) and creates semantic representations (III) that are conveyed to node syntax (II) which in turn is linearized (I) in actual speech.

Words, that is, root lexemes (L) and their modifiers, or morphemes (M), have phonetic, syntactic, semantic, phenomenological (connotative), and epistemico-pragmatic properties, and can therefore be seen as related to all instances, while being different from them all; they are linguistic "signs", in the classical view, not structures. Language seems to be the evolutionary result of an encounter of two distinct mental phenomena: expressive signs and structural representations. We could have had signs – such as names, words to call on or point to things – and still not have language. We could have had structures – such as: gestures to show states of affairs – and still not have language. Now, we happen to have an integrated

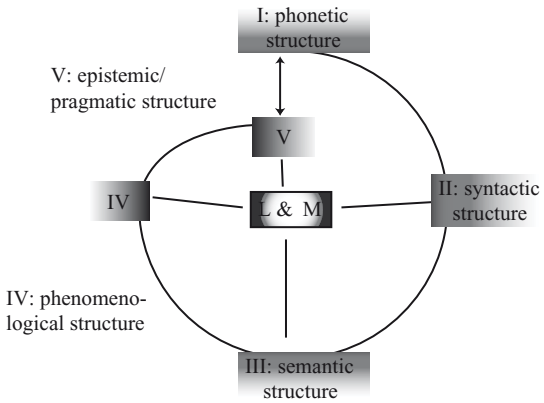

Fig. 1 Five levels of processing without a "deep structure"

combination of these two: language. I suspect that such an evolutionary event of integration is the story behind the existence of two separated and connected linguistic cortex areas, Broca's and Wernicke's.

We can of course think without speaking, and then still use words, that is, terms, in our inner processing, including referential terms in diagrams and equations. In music, inversely, we can easily imagine and perform syntactically structured phrases without words.

4 Language and Culture

We "think" when our mind mobilizes internal or external resourses to make sense of, or develop, or find ways around, or solve, a present "problem". The outcome of thinking is some sort of solution or conclusion or understanding that makes us either modify our behavior, initiate some action, or just pass on to other, possibly related "problems". Again, a "problem" is, I suggest, essentially a practical or theoretical situation that inhibits the flow of our doings or our thinking and calls for attention and particular treatment in order to reopen the flow of our doings.

Through the levels of language, we communicate and either share or reject each others' thinking (in the broad sense indicated). In the perspective of our cultural evolution, the differentiation of languages might suggest that rejection is the more common outcome; the multitude and diversity of mutually opaque languages in cultures that occupied fertile territories on all continents, before modern civilizations imposed their semiotic regimes, allow us to think that incommunication can be a peace condition, whereas communication within a shared language can lead to social fractures, conflicts, and violence – which may be why religion seems to become important for social coherence within a language community. The paleolithic spreading of our species might be due to the same factors that motivate contemporary migrations, tragical outcomes of mainly intralingual conflicts. We think, therefore we speak, and therefore we fight; speech acts (such as declarative provocations) are core factors in the eruption of human violence.[14] Religion damps down the effects of thinking and speaking; but in large populations sharing language, any religion invariably breaks into conflicting beliefs and sects.

We think, therefore we gesture and exchange signs of all kinds; those are, as mentioned initially, more immediate and direct but more laconic expressions of thinking. Symbols in social life are predominantly deontic signals that regulate

[14]Neigboring tribal cultures that share language fight far more violently than neighboring tribes that speak different languages. In relatively densely populated areas in Africa, the linguistic diversity is considerable. Here is an example: "Cameroon is home to 230 languages. These include 55 Afro-Asiatic languages, two Nilo-Saharan languages, and 174 Niger-Congo languages. This latter group is divided into one West Atlantic language (Fulfulde), 32 Adamawa-Ubangui languages, and 142 Benue-Congo languages (130 of which are Bantu languages)". From a Wikipedia article.

conventional behaviors; they are semantically stabilized through language. Rituals are frozen behaviors that in turn regulate the use of language, mostly by ruling out unwanted discourse. In religious behavior, minimizing local conflicts, symbols and rituals must therefore be harmonized, and language must therefore be strictly controlled. In secular behavior, symbols and rituals often compete and creatively contradict each other – as in political life, and in the juridical domain. The most intriguing signs are the icons: their relation to thinking is so direct that intuitive and still highly abstract communication through images is an everyday phenomenon.[15] Pictorial art may be the iconic practice that challenges human semio-linguistics more than any other semiotic practice; "good" art seems invariably characterized by the fact of defying semantic stabilization by language. We find a work of art masterful if it makes us feel that is means more than any commentary; it then beats language, so to speak, and still it relates directly to thinking. It thus makes thinking feel unending. Unending thinking in turn leads to our "infinite" issues, where thinking and feeling become indistinguishable.

Paradoxically, literature can do what art does against language, but through language. Poetry, fiction, theater, all literary forms of art, are in fact iconic uses of language. In literature, language induces thinking as *icons of* thinking, and the latter still feels unending. It may be reasonable to say that language could have "killed" thinking, by locking it into its syntactic structures and their semantic frames and domains; but that literature, developed out of art, then saved it from such a sad ending, and handed it over to science and life in general.

References

Brandt PA (2004) Spaces, domains, and meaning. Essays in cognitive semiotics. Peter Lang Verlag, Bern European Semiotics Series, No 4

Chafe W (2005) The relation of grammar to thought. In: Butler, Gómez-González, Doval-Suárez (eds) The dynamics of language use: functional and contrastive perspectives, John Benjamins, Amsterdam pp 57–78

Croft W (2007) The origins of grammar in the verbalization of experience. Cogn Linguist 18–3:339–382

Fillmore CJ, Atkins BT (1992) Towards a frame-based organization of the lexicon. In: Adrienne Lehrer, Eva Kittay (eds) Frames, fields, and contrasts: new essays in semantics and lexical organization. Lawrence Erlbaum, Hillsdale, pp 75–102

Hagoort P, van Berkum J (2007) Beyond the sentence given. Philos Trans Ro Soc 362:801–811

Lotman Y, (1990) (Eng. transl. Ann Shukman) The universe of the mind. A semiotic theory of culture. Indiana University Press, Bloomington

Merleau-Ponty M (1960) Signes. Gallimard, Paris

Schank RC, Abelson RP (1977) Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum, Hillsdale

[15]The Danish cartoon crisis (triggered by the drawing of a muslim with turban-bomb) is a striking example; though it remains unclear what precisely the cartoon is "saying", its provocative force is empirically proven.

Thinking and Emotion: Affective Modulation of Cognitive Processing Modes

Annette Bolte and Thomas Goschke

Abstract In this chapter, we review empirical findings showing that positive and negative affective states are accompanied by qualitatively different information-processing modes. Specifically, positive moods and emotions appear to be associated with a more flexible processing mode as indicated by a broadened scope of attention, activation of weak or unusual associations, and facilitated switching between cognitive sets. We interpret these findings within a general theoretical framework according to which different modes of thinking serve complementary or even antagonistic adaptive functions in the planning and control of goal-directed action. In contrast to the widespread view that positive affect has exclusively beneficial consequences such as increased creativity and flexibility, we argue that different emotions and moods and the processing modes associated with them incur complementary costs and benefits. Thus, consistent with recent findings, positive and negative affect have advantages and disadvantages depending on the processing requirements of the to-be-performed task.

A. Bolte (✉)
Department of Psychology, Technische Universität Dresden, 01062, Dresden, Germany
e-mail: bolte@psychologie.tu-dresden.de

T. Goschke
Department of Psychology, Technische Universität Dresden, 01062, Dresden, Germany
e-mail: goschke@psychologie.tu-dresden.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_18, © Springer-Verlag Berlin Heidelberg 2010

1 Introduction

Our thinking is deeply influenced by our emotions and moods.[1] For instance, it is a familiar experience that in a positive mood one has the impression that one's life is full of opportunities; one comes up with a multitude of new ideas; and when thinking of the past primarily all the good things one has experienced come to mind. In contrast, in a sad or depressed mood one's attention appears to be focused in a rigid and narrow manner on a single negative topic, and one's mind is occupied with ruminations about past failures, negative experiences, or dismal prospects. As these examples illustrate, different moods – and in general emotional states – can influence thought processes in two principle ways: First, emotions and moods influence the contents of thought, i.e., *what* we focus our attention on, retrieve from memory, and reflect upon. Secondly, moods and emotions modulate the mode of thought, i.e., *how* we think, decide, and process information. As regards the first type of affective influence, since Bower's (1981) influential article on mood-congruency effects, a large empirical literature has accumulated showing that positive or negative moods facilitate the encoding and retrieval of memory contents with an affective valence that is congruent with one's current mood (although exceptions to this general effect have been reported and a number of boundary conditions for mood-congruency effects have since been identified; for reviews see Blaney 1986; Goschke 1996).

In this chapter, we focus instead on the second type of affective influences on thought processes, that is, we provide a selective review of evidence indicating that moods and emotions are associated with qualitatively different modes of thought and information-processing. Although the affective modulation of different modes of cognitive processing has been investigated less extensively than content-specific (e.g., mood-congruency) effects, there is now convincing evidence that emotions and moods do systematically influence the way we think and process information in various domains, including creative problem-solving, activation of semantic associations, selective attention, and cognitive control (for reviews see Ashby et al. 1999; Davidson et al. 2000; Erk and Walter 2000; Fiedler 2001; Forgas 2000; Fredrickson 2001; Isen 1999, 2004; Kuhl 1983, 2000; Martin and Clore 2001).

[1] In this chapter, we use the term *emotion* to denote psycho-physiological response patterns, which rest on more or less complex evaluations of events or actions in the light of an organism's needs, motives, and goals; which are accompanied by changes in the peripheral nervous system (e.g., increased arousal); which are controlled by specific brain circuits (e.g., the amygdale in the case of fear); which motivate the organisms toward particular categories of action (e.g., fight or flight); which are often accompanied by specific facial and postural expressions; and which are usually (but not always) associated with a specific qualitative subjective experience. In contrast, we use the term *mood* to denote more enduring, mild emotional states, which are not necessarily directed towards or caused by a particular object or event, need not be in the focus of attention, and have a non-focal ("colorizing") experiential quality. Finally, we will use the terms *affect* or *affective state* as generic summary terms subsuming both moods and emotions in the more narrow sense as defined above.

Our review will be guided by the theoretical idea that different modes of thinking (e.g. analytic vs. holistic; explicit vs. intuitive; focused vs. divergent) serve different and sometimes antagonistic adaptive functions in the planning and control of goal-directed action. As a direct consequence of this assumption, we assume that different emotions and moods and the processing modes associated with them incur complementary costs and benefits. Thus, in contrast to the widespread view that positive affect has primarily beneficial consequences like, for instance, increased creativity and cognitive flexibility, we assume that positive affect (like any emotional state) can have both advantages and disadvantages, depending on the processing requirements of the to-be-performed task.

2 Antagonistic Adaptive Functions and Complementary Modes of Thinking

Our central assumption that different modes of thinking serve complementary adaptive functions derives from a general theoretical framework, according to which organisms in a changing and uncertain world face antagonistic adaptive challenges or "*control dilemmas*" (Dreisbach and Goschke 2004; Goschke 1996, 2000, 2003). These control dilemmas afford a dynamic and context-sensitive balance between complementary modes of thought and action.

As an example, consider the *selection-monitoring dilemma*: On the one hand, agents must focus their attention selectively on task-relevant information and suppress distracting stimuli in order to prevent crosstalk and interference. On the other hand, however, it is equally important for an agent to monitor the environment for potentially significant information, even if this information is not relevant for the ongoing task. It would hardly be adaptive to focus attention so exclusively on a current goal (e.g., writing a chapter on emotion and thinking) that task-irrelevant information (e.g., the smell of smoke indicating a fire) is ignored completely. Thus, focusing attention vs. monitoring for potentially significant information incur complementary benefits and costs: while a narrow scope of attention and the suppression of distracting information reduces interference, it increases the risk to oversee potentially significant stimuli. Conversely, whereas a less focused, more distributed scope of attention increases the likelihood to detect potentially significant stimuli, it incurs a cost in terms of increased distractibility (cf. Dreisbach and Goschke 2004).

A second example is the *exploration–exploitation dilemma*: On the one hand, it is adaptive for organisms to rely on well-established routines and to select actions on the basis of acquired knowledge about reward contingencies and action outcomes. On the other hand, it is equally important to explore novel actions, which may lead to yet unknown, but potentially even better outcomes (cf. Aston-John and Cohen 2005; Doya 2008; Goschke 1996; Sutton and Barto 1998). Exploitation and exploration incur complementary benefits and costs: Whereas an exclusive reliance on the exploitation of learnt contingencies prevents an organism from finding even

more rewarding actions, an excessive tendency to explore novel actions leads to erratic behavior that is insufficiently constrained by prior experience. In more general terms, this dilemma can also be phrased in terms of the question how an agent decides in the face of uncertainty whether to interpret novel experiences in the light of preexisting beliefs (assimilation), or whether to revise beliefs in the light of new experiences (accommodation) (Piaget 1975).

To the degree that emotions and moods alter the balance between complementary modes of thinking and information-processing, a particular emotional state should likewise be associated with complementary costs and benefits. Thus any affective state (and the associated processing mode) should lead to advantages in certain tasks, but at the same time impair performance in other tasks, depending on whether the processing requirements of the task fit with the prevailing processing mode promoted by the current emotion.

3 Theoretical Views on the Affective Modulation of Cognitive Processes

Several theories of the affective modulation of cognition contain the assumption that positive and negative moods or emotions are associated with qualitatively different processing styles. In particular, most theories agree that a positive mood is associated with a more flexible processing style, that is characterized by the activation of widespread associative relations, a broadened focus of attention, and an increased readiness to explore new ideas and opportunities for alternative thoughts and actions (e.g., Bolte, Goschke, & Kuhl 2003; Dreisbach and Goschke 2004; Fiedler 2001; Fredrickson 2001; Fredrickson and Joiner 2002; Goschke 1996; Isen 1999; Kuhl 2000). For instance, Alice Isen has pursued an extended research program on the effects of positive mood on cognition that rests on the hypothesis that positive compared to neutral or negative affect promotes a more flexible cognitive organization (for reviews see Isen 1999, 2004). In a similar vein, Fredrickson (2001) has developed a "broaden-and-build theory" according to which positive emotions or moods are associated with an expanded focus of attention and an increased tendency to explore new ideas or action alternatives, which renders cognitive processing more flexible, explorative, and creative. In contrast, negative emotions or moods are assumed to induce a narrow focusing of attention and restrict the variety of alternative thoughts that come to mind. According to Fredrickson, negative emotions evolved in order to support specific-action tendencies and to prepare the organism for adaptive action (for example, attack in the case of anger, or escape in the case of fear). In contrast, the adaptive function of positive emotions like joy or contentment is not seen in promoting specific behavioral tendencies, but rather in expanding the repertoire of thought and possible actions.

Fiedler (2001), in his theory of mood-cognition interactions, also distinguishes between two complementary information-processing functions: *information conservation* and *active generation*. The first function consists in the encoding and

conservation of given input information, whereas the second function consists in the active generation of new information based on prior knowledge, as, for instance, in the development of new ideas during creative thinking. Fiedler distinguishes between different cognitive sets in appetitive and aversive situations: whereas appetitive settings encourage exploration, creativity, and the generation of new ideas, in aversive settings the organism must be attentive to potentially threatening stimuli and avoid making mistakes. Therefore a negative mood (that is typically associated with aversive settings) is assumed to support the conservative function (i.e., focusing on stimulus details or factual information), whereas a positive mood supports active generation (i.e., making new inferences and engaging in creative thought).

From a related perspective, Bless (2001) proposed a mood-and general-knowledge model, according to which individuals in benign situations rely on their general knowledge structures, whereas in problematic situations attention is focused on specific details. Bless assumes that these different information-processing modes are adaptive to the individual. Because problematic situations are usually deviations from routine situations individuals would be poorly advised to rely on the knowledge they usually apply and should better focus on the specifics of the current situation. In contrast, in benign situations less attention must be paid to specific details. Here it is adaptive to save processing resources allocated to the specifics of the situation and to direct resources toward other tasks or to generate and test new ideas (see also Schwarz 2001).

A particularly elaborated theory of the affective modulation of complementary cognitive systems and processing modes has been developed by Kuhl (2001). A central assumption of Kuhl's personality systems interactions theory ("PSI theory") consists in several affect modulation hypotheses, which specify how affective states modulate the relative balance and interaction between cognitive systems, which mediate qualitatively different aspects of information-processing and action control. More specifically, Kuhl assumes that an increase in negative affect promotes an analytic processing mode in which attention is focused on isolated details of the current stimulus situation, whereas the down-regulation of negative affect promotes a holistic processing mode, that is characterized by access to stored experiential knowledge (including personal preferences and implicit motives) and facilitates the integration of isolated pieces of information into a widespread network of coherent memory representations.

4 Selective Review of Evidence for the Affective Modulation of Complementary Modes of Thinking

Despite differences in specific assumptions, most of the mentioned theories agree that positive affective states are accompanied by increased cognitive flexibility as indicated by a broadened scope of attention, the activation of weak or unusual associations, and an increased readiness to explore new thoughts and actions. This assumption is supported by a growing body of empirical evidence on effects of

positive and negative affect on the mode of information-processing. In this section we selectively review studies on the affective modulation of thought processes in four domains:

(1) Creative problem-solving and generative thought
(2) Activation of semantic associations
(3) Selective attention
(4) Cognitive control

4.1 Affective Modulation of Creative Problem Solving and Generative Thought

In an early study, Isen et al. (1987) investigated the effects of experimentally induced mood states on creative problem solving. One of the tasks they used was Duncker's (1945) candle task, in which participants received a box of matches, pushpins, and a candle. The task was to fix the candle on the wall with these implements such that no wax would drop on the floor (the solution is to empty the box of matches and pin it on the wall with the pushpins, such that it serves as a candleholder). Participants, in whom a positive mood had been induced by seeing a few minutes of a comedy film, came up with the solution within the specified time almost four times as often as participants in a negative or neutral mood.

More recent studies have obtained similar results and provide converging evidence that a positive mood reduces functional fixedness in solving insight problems. For instance, Gasper (2003) had participants in happy, sad, and neutral moods perform a classic problem-solving task designed to investigate the perseveration and breaking of mental sets (Luchins 1942). In this task, participants initially complete a series of similar problems, which can be solved only by using a relatively complex strategy and which serve to establish a stable mental set. In a subsequent phase of the experiment, problems are presented that can be solved by using either the established strategy or an obvious and much simpler strategy. In Gasper's experiment, participants in a sad mood relied on the established mental set until they received feedback that their strategy may be problematic, whereas participants in a positive mood were more likely to abandon the established mental set on their own.

As already mentioned, Fredrickson (2001) in her broaden-and-build theory assumes that positive affect increases the tendency to explore new ideas or action alternatives. Consistent with this assumption, Fredrickson and Branigan (2005) found that a positive mood enlarged the thought-action repertoires that participants generated. After participants had watched emotion-eliciting videos, they were asked to indicate all of the action urges they had at that moment. Participants in a positive mood generated a greater number of action-urges than people in a negative or neutral mood.

Consistent with this finding, it has been found that participants in a positive mood (induced by having participants read funny short stories) performed better

than participants in a neutral mood on a fluency test, in which they had to produce as many novel uses for a given object (e.g., a cup) as possible (Phillips et al. 2002, Experiment 2). There was no reliable effect of positive mood on a superficially similar letter fluency task (producing as many words beginning with the letter A), although a correlational analysis revealed that the more positively participants rated their mood at the end of the experiment, the more words did they produce in the letter fluency task.

In summary, the studies reviewed in this section indicate that a positive mood improves performance in creative problem solving tasks by reducing functional fixedness, facilitating the breaking of mental sets, and by enlarging the thought-action repertoires that participants generate.

4.2 Affective Modulation of Semantic Associations

Several studies have examined the effects of positive mood on the activation and retrieval of semantic associations. In an early study, Isen et al. (1985) showed that participants in a positive mood gave more unusual first-associates to neutral words than did subjects in neutral mood or subjects who received no mood induction. In a further experiment, in which word type (positive, neutral, or negative) was a second independent variable along with induced mood, participants produced associates to positive words that were more unusual and diverse than were those to other words. Thus positive mood as well as processing words with a positive emotional valence promoted the activation and retrieval of more unusual associates.

In a further study, Isen et al. (1987) used the Remote Associates Test (Mednick and Mednick 1967), in which participants are presented three clue words and have to find a common associate (e.g., *mower* – *atomic* – *foreign* with the associated solution word *power*). Participants in whom a positive mood had been induced came up with more solution words compared to participants in a neutral mood. A condition in which a negative mood was induced and two additional control conditions in which participants engaged in physical exercise (intended to increase unspecific arousal) failed to produce comparable improvements in creative performance. This finding was recently replicated by Rowe et al. (2007), who also found that participants in a positive mood (induced by listening to pieces of music) performed reliably better in the Remote Associates Test than participants in a negative or neutral mood. The authors interpreted this as evidence that a positive affect induces a more open and exploratory mode of attention to both internal and external sources of information and thereby facilitates access to distantly related associates in memory.

Bolte et al. (2003) showed in addition that a positive (happy) mood facilitates not only the explicit retrieval of remote associates from memory, but also improves the ability to make *intuitive* judgments about the semantic coherence of word triads, even if participants do not consciously retrieve the associated solution word. Participants were presented three clue-words which were either coherent in the

sense that all three words were weakly associated with a common fourth concept, or which were incoherent, i.e., there was no common associate. Participants in a neutral mood discriminated coherent and incoherent triads reliably better than chance level even if they did not consciously retrieve the solution word. The induction of a positive mood reliably improved intuitive coherence judgments, whereas participants in a negative mood performed at chance level. The authors concluded that positive mood potentiates spread of activation from the three clue words to the remote associate in memory, which gives rise to an intuitive impression of semantic coherence, even if the common associate is not accessible to consciousness. By contrast, a negative (sad) mood obviously restricted the spread of activation to close associates and dominant word meanings, thereby impairing the ability to intuitively judge the semantic coherence of the word triads. Consistent with this interpretation, Baumann and Kuhl (2002) found that individuals with a reduced ability to down-regulate negative affect (so-called state-oriented individuals) performed worse on the intuitive coherence task when being in a sad mood, compared to participants who tend to down-regulate negative emotions (so-called action-oriented individuals).

Further evidence that positive affect facilitates the processing of distantly related semantic concepts stems from an event-related potential (ERP) study by Federmeier et al. (2001). Participants read sentences which ended either with a highly expected word, an unexpected word from an expected semantic category, or an unexpected word from a different category. For participants in a neutral mood, amplitudes of the N400 component of the ERP, which is sensitive to the degree of semantic deviations (Federmeier and Kutas 1999) were smallest for expected items from an expected category. Moreover, N400 amplitudes were smaller for unexpected items from an expected category than for words from unexpected category. By contrast, for participants in a positive mood N400 amplitudes did not differ in response to the two types of unexpected items. The authors concluded that a positive mood facilitated the processing of unexpected, but distantly related items.

This conclusion fits with an earlier result of Isen and Daubman (1984) who had participants classify exemplars, which differed in their typicality with respect to a given category, as members of the category. Participants in a positive mood, induced by means of seeing a few minutes of a comedy film, showed a stronger tendency to classify less typical exemplars as members of a category than participants who saw a short film about mathematics (neutral mood) or a film about concentration camps (negative mood). In two further studies, in which participants had to group different exemplars to categories, participants in positive mood used fewer categories to sort the exemplars than participants in neutral or negative mood, indicating that a positive mood promotes an over-inclusive mode of categorization.

In conclusion, there is converging evidence that a positive mood facilitates access to weak or remote associates in memory, improves intuitive judgments in a semantic coherence task, and increases the tendency to view distant or untypical exemplars as members of a semantic category. Although different processes may underlie the effects of positive mood in the different tasks discussed so far, it is a plausible hypothesis that many of the discussed findings reflect a common

underlying mechanism, namely the activation of widespread associative networks including weak, unusual, or remote associates. That is, the fact that positive mood induces a broadened scope of semantic associations may not only account for improved performance in tasks requiring activation or retrieval of remote associates, but also for the increased rate of solutions in insight problems, which require one to break habitual cognitive sets and to recognize novel or unusual relations between familiar elements. By contrast, a negative (especially a sad) mood appears to induce a more analytic and focused style of processing, which is characterized by a more restricted spread of activation to close associates in memory, thereby impairing performance on remote associate tasks as well as on insight problems.

4.3 Affective Modulation of the Scope of Selective Attention

The studies reviewed so far may give the impression that a positive mood has in most cases beneficial consequences like increased creativity and cognitive flexibility. However, in the introduction we proposed that different modes of thinking (e.g., focused vs. divergent) serve complementary adaptive functions and that different emotions and moods, by inducing qualitatively different processing modes, will have both advantages and disadvantages, depending on the processing requirements of to-be-performed task. More specifically, we expect that positive affect should impair performance when a more focused style of processing or a narrow scope of attention is required by a given task. In this and the next section we review recent studies on the affective modulation of selective attention and cognitive control processes, which indicate that positive affect can also incur a performance cost in tasks requiring focused attention or inhibition of distracting information.

It has long been assumed that positive affect is associated with a broadening of the scope of attention, whereas negative affect (especially when associated with high arousal) leads to a narrowed focus of attention (e.g., Easterbrook 1959). This assumption fits with everyday observations like the so-called "Weapon focus," which denotes the fact that victims of a violent assault often show an improved memory for central details of the event (e.g., the weapon used), but cannot remember other aspects of the experience, indicating an extreme narrowing of the focus of attention. Consistent with these observations, laboratory experiments have shown that negative affect elicited by aversive pictures (e.g., of a traffic accident) induces a narrow focus of attention on central details of the scene at the cost of peripheral aspects (e.g., Christianson and Loftus 1991).

On the other hand, there is evidence that positive affect has the opposite effect of broadening the focus of attention. For instance, in some studies the effects of positive mood on the processing of global or holistic vs. analytic details of visual stimuli were examined. Gasper and Clore (2002) asked participants to indicate which of two stimuli was more similar to a reference stimulus. Stimuli resembled each other either with respect to their global shape or with respect to the local elements from which they were composed. For instance, a large square made

out of four small triangles resembled a large square made out of small squares with respect to global shape, whereas a large triangle made out of small triangles resembled a large square made out of small triangles on the level of local elements. As predicted, individuals in a sad mood were less likely than those in a positive mood to classify figures on the basis of global features. Likewise, Fredrickson and Branigan (2005) reported that participants, in whom a positive mood had been induced by having them view video clips eliciting amusement or contentment, showed a stronger bias than participants in neutral or negative moods to judge the similarity between visual stimuli on the basis of their global resemblance.

Direct evidence that a broadened scope of attention induced by positive affect may incur performance costs due to reduced attentional selectivity has recently been reported by Rowe et al. (2007). They induced a happy or a sad mood by having participants listen to selected pieces of music, whereas a neutral mood was induced by having participants read factual statements about Canada. After the mood induction, participants performed the Remote Associates Test (described above) and a visual selective attention task, the Eriksen flanker task (Eriksen and Eriksen 1974). In the flanker task participants had to respond to a central target stimulus (e.g., a letter) and ignore irrelevant flanking distracters. When target and distracters are mapped to incompatible responses, responses are usually slowed, indicating a failure of selective attention and an unwanted influence of the to-be-ignored flankers. As expected, participants in a positive mood showed a greater flanker interference effect relative to participants in sad or neutral moods, which was due to a disproportionate slowing to incompatible flankers under positive mood. Thus positive mood obviously broadened the scope of visuospatial attention, thereby increasing the (unwanted) processing of distracting flanker stimuli. Interestingly, as already reported above, in this study positive mood improved participants' performance in the Remote Associates Test, and within the positive mood group a significant correlation was found between the slowed response times associated with flanker incompatibility and the number of correctly identified remote associates. This indicates that enhanced access to remote associates was correlated with impaired visual selective attention and nicely demonstrates within one study, that the broadened scope of attention induced by positive mood can incur both costs and benefits, depending on the particular task requirements.

4.4 Affective Modulation of Cognitive Control

Further evidence for complementary costs and benefits of positive affect stems from recent studies of the affective modulation of cognitive control processes. Cognitive control is a summary term denoting mechanisms which serve the maintenance of goals in the face of distraction, the inhibition of task-irrelevant stimuli or prepotent, but inadequate responses, and the flexible reconfiguration of behavioral dispositions in response to changing goals or task demands (cf. Goschke 2003, 2007; Miller and Cohen 2001).

Initial studies of the affective modulation of cognitive control yielded mixed results. While some researchers found that a positive mood impaired performance on tasks requiring executive control like the Tower of London task (Oaksford et al. 1996) or when participants had to switch between different tasks (Phillips et al. 2002), others found that phasic increases in positive affect induced by positive emotional words reduced interference in the Stroop task (Kuhl and Kazén 1999). Moreover, Gray (2001) found that affective states either impaired or improved performance in a working memory task depending on whether the task involved spatial or verbal information. These inconsistencies may at least in part be due to the fact that different "executive" tasks involve different or even antagonistic requirements, as, for instance, the maintenance and shielding of a current task-set versus the flexible switching between task-sets. As we noted above, whether positive affect impairs or improves performance in tasks requiring cognitive control should critically depend on the specific control demands imposed by a particular task.

This hypothesis was directly tested by Dreisbach and Goschke (2004) who investigated differential effects of positive affect on complementary cognitive control functions. Specifically, these authors predicted that positive affect reduces perseveration and facilitates flexible switching of cognitive sets, but at the same time incurs a cost due to increased distractibility. To test this hypothesis, the authors used a task in which participants were first trained to respond to target stimuli appearing in a prespecified color (e.g., red), while ignoring distracter stimuli in a different color (e.g., green). After a block of 40 trials, participants were transferred to one of two switch conditions. In one condition, after the switch they had to respond to stimuli in a new color that had not appeared before (e.g., blue), while all the distracters appeared in the previous target color (i.e., red). In this condition, increased flexibility should facilitate switching to novel stimuli, as indicated by *decreased* switch costs (the authors termed this condition the *perseveration condition*, because switch costs primarily reflect the degree to which the previously relevant cognitive set perseverates). In the second switch condition, participants had to respond to stimuli in the previously to-be-ignored color (i.e., green), while all the distracters appeared in a new color (i.e., blue). In this condition, increased flexibility or a broadened scope of attention should bias participants' attention toward the novel distracters, thereby producing *increased* switch costs (one may term this condition the *distractibility condition*, because switch costs reflect the tendency to focus attention on novel distracters). To induce phasic affective responses, in different blocks either positive, neutral, or negative affective pictures from the International Affective Picture System (IAPS), Lang et al. (1998) were presented briefly before each imperative stimulus. Consistent with the authors' predictions, the presentation of positive pictures had opposite effects in the two switching conditions: Whereas the presentation of positive affective pictures almost completely *eliminated* the switch cost in the perseveration condition, it reliably *increased* the switch cost in the distractibility condition. This dissociation is consistent with the hypothesis that phasic increases in positive affect increase cognitive flexibility, albeit at the cost of increased distractibility. Importantly, the pattern of findings could not be accounted for by the higher arousal potential of the positive compared to neutral pictures,

because negative emotional pictures which matched the arousal potential of positive pictures did not differ in their effects from neutral pictures.

5 Conclusions and Open Questions

The findings reviewed in this chapter show that induced tonic moods as well as phasic emotions exert strong influences on the prevailing mode of cognitive processing. In particular, in line with the theories outlined in the introduction, there is converging evidence that positive affect is associated with a more flexible processing style that is characterized by the activation of widespread networks of weak or remote associates in memory, a broadened scope of attention, and an increased readiness to explore new ideas and opportunities for alternative actions. The mode of thinking induced by positive affect usually improves performance in tasks requiring a more global or "holistic" style of information processing, such as the remote associates task (requiring the activation of widespread associative networks), insight problems (requiring the breaking of habitual cognitive sets and the detection of novel or unusual relations among cognitive elements), or fluency tasks (requiring the generation of a wide variety of alternative action options). However, as predicted by our complementary processing modes framework, positive affect can also incur performance costs in selective attention and cognitive control tasks, when the tasks require a more focused style of processing or the inhibition of distracting sources of information.

This conclusion fits with the general assumption outlined in the introduction, that different processing modes serve complementary adaptive function in the control of action. Accordingly, many theorists implicitly or explicitly rely on evolutionary considerations when justifying specific hypotheses concerning the affective modulation of cognitive processes. As described in the introduction, it is frequently assumed that negative emotions evolved to prepare the organism for specific adaptive action in response to danger, threat, or challenged goal pursuit. Accordingly, it appears plausible to assume that negative emotional states are associated with a "conservative" mode of processing (Fiedler 2001) and a focusing on details of potentially threatening stimuli (Bless 2001; Fredrickson 2001). Conversely, positive emotions are usually interpreted as signals that goal pursuit runs smoothly and there are no immediate or anticipated threats that one must cope with. Thus it appears plausible that positive emotional states promote an exploratory mode characterized by creative thought, new inferences, and the generation of unusual ideas, which may serve to expand the repertoire of thoughts and possible actions (Fiedler 2001; Fredrickson 2001; Goschke 1996; Kuhl 1983, 2001). However, as is the case for most evolutionary accounts, one should be aware of the fact, that such hypotheses are plausible post-hoc accounts of the adaptive function of mental processing modes, which may be difficult to test in a strict experimental sense.

Another influential interpretation for affective influences on cognitive processing modes (that is not incompatible with an evolutionary-functional analysis) rests on

the assumption that emotions and moods serve an informational function in that they indicate whether a situation is benign or problematic, thereby tuning cognitive strategies to meet the respective situational requirements (e.g., Schwarz 2001). For instance, if positive moods or emotions signal the absence of problems or obstacles, it may inform the organism that there is no risk in engaging in a more intuitive, exploratory, or creative mode of thought. By contrast, negative moods or emotions usually indicate the presence of conflicts, problems, or dangers, and may thus signal that a more analytic problem-solving mode of processing is required. Thus effects of moods and emotions on, for instance, the activation of remote associates or creative problem-solving may reflect the informational function of affective states. It is an open question, however, whether the informational content of affective states in general or moods in particular influences cognitive processes primarily by way of a deliberate strategy change, or whether moods and emotions can also modulate cognitive processing modes in a more automatic way. Likewise, it is an open question whether moods and emotions influence cognitive processes or judgments only when they are experienced as relevant sources of information in an ongoing task (e.g., Schwarz and Clore 1983; Schwarz 1990; Schwarz and Bless 1991; Schwarz et al. 1991), or whether they influence cognitive processing modes also in a more direct way, that is, independently from whether or not a current mood or emotion is attributed to a specific cognitive source.

In closing our review we would like to point to three further open questions for future research. First, in this chapter we have neglected differences between specific emotions (e.g., anger, fear, disgust, sadness). Different emotions most likely developed as evolutionary answers to specific adaptive challenges (Panksepp 1982; LeDoux 1996) and it is therefore very likely that different emotions like fear and anger – even though they may share a similar valence – are associated with qualitatively different processing modes (Dörner 1999). Thus to speak simply of positive and negative emotions or moods is clearly an oversimplification. Closely related to this point is the requirement to distinguish more systematically between the effects of tonic moods and phasic emotions. Although in most of the studies reviewed here, positive mood had similar effects as the induction of phasic increases in positive emotions (even if such phasic emotional responses were not accompanied by enduring mood changes), it is an important empirical question under which conditions moods and emotions differ in their effects on cognitive processes.

A second question concerns the observation that in many studies using tonic mood induction techniques in nonclinical subject populations, induced positive moods had stronger effects on cognitive processes than negative moods. One possible explanation is that negative mood induction methods (e.g., listening to sad music) may not result in equally strong or unambiguous mood changes, or that participants use “mood-repair” strategies to counter-regulate negative affective states (Isen et al. 1987). Moreover, many of the cognitive tasks typically used to examine effects of mood on cognition are relatively demanding and boring for the participants and will thus often elicit a negative change of participants’ mood independently from the intended mood induction.

A third – and theoretically most important – question is which psychological processes and neurobiological mechanisms underlie the effects of emotions and moods on cognitive processing modes. At present, relatively little is known about the specific mechanisms by which affective states exert their modulating influence on cognitive processing modes. On a psychological level, one promising hypothesis holds that affective states influence the settings of global *parameters* of the cognitive system, which regulate the *mode* of information-processing independently from the processed contents (Dörner 1999; Doya 2008; Erk and Walter 2000). An example of such a global processing parameter is the *signal-to-noise ratio*, which regulates the degree to which the cognitive system engages in exploratory behavior; another example is the *scope of attention*, which regulates the degree to which attention is focused or distributed; a third example is the *switching or updating threshold*, which regulates how easily the current content of working memory is updated vs. how strongly this content is shielded from distraction. It is an interesting hypothesis for future research that different affective states are associated with specific pattern or configurations of such processing parameters (for an elaborate version of this hypothesis and computational models based on it see Dörner 1999; Dörner et al. 2002). On a neurobiological level, there is evidence that some of the effects of affective states on the setting of global processing parameters may be mediated by the action of specific neuromodulatory systems. For instance, Ashby et al. (2002, 1999) have hypothesized that some of the effects of positive affect on cognitive processes are mediated by increased levels of brain dopamine (DA) in frontal cortical areas, notably the anterior cingulate cortex (ACC). Increased DA levels in the ACC are assumed to enhance the ability to overcome dominant responses and to facilitate flexible switching of cognitive sets. There is indeed increasing evidence from neurobiological and neuroimaging research that neuromodulators like dopamine, serotonin, and norephinephrine influence prefrontal cortical functions involved in thinking and planning, the maintenance vs. updating of information in working memory, and the regulation of focused vs. distributed modes of attention (e.g., Braver and Cohen 2000; Cools 2008; Dreisbach et al. 2005; Müller et al. 2007; Roberts 2008). While a discussion of these findings is beyond the scope of this chapter, it will be a major goal for future research to relate behavioral studies on the affective modulation of cognitive processing modes more closely to underlying neuromodulatory systems and their effects of neural systems involved in thinking and cognitive control.

References

Ashby FG, Isen AM, Turken AU (1999) A neuropsychological theory of positive affect and its influence on cognition. Psychol Rev 106:529–550

Ashby FG, Valentin VV, Turken AU (2002) The effects of positive affect and arousal on working memory and executive attention: Neurobiology and computational models. In: Moore S, Oaksford M (eds) Emotional cognition: from brain to behaviour. John Benjamins, Amsterdam, pp 245–287

Aston-John G, Cohen JD (2005) An integrative theory of locus coeruleus-norephinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci 28:403–450
Baumann N, Kuhl J (2002) Intuition, affect, and personality: unconscious coherence judgments and self-regulation of negative affect. J Pers Soc Psychol 83:1213–1223
Blaney PH (1986) Affect and memory: a review. Psychol Bull 99:229–246
Bless H (2001) The relation between mood and the use of general knowledge structures. In: Martin LL, Clore GL (eds) Mood and social cognition: contrasting theories. Lawrence Erlbaum Associates, Mahwah, NJ, pp 9–29
Bolte A, Goschke T, Kuhl J (2003) Emotion and intuition: effects of positive and negative mood on implicit judgments of semantic coherence. Psychol Sci 14:416–421
Bower GH (1981) Mood and memory. Am Psychol 36:129–148
Braver TS, Cohen JD (2000) On the control of control. The role of dopamine in regulating prefrontal function and working memory. In: Monsell S, Driver J (eds) Control of cognitive processes: attention and performance XVIII. MIT Press, Cambridge, MA
Christianson S-A, Loftus EF (1991) Remembering emotional events: the fate of detailed information. Cogn Emot 5:81–108
Cools R (2008) Dopaminergic modulation of flexible cognitive control: the role of the striatum. In: Bunge SA, Wallis JD (eds) Neuroscience of rule-guided behaviour. Oxford University Press, Oxford, pp 313–334
Davidson RJ, Jackson DC, Kalin NH (2000) Emotion, plasticity, context and regulation: perspectives from affective neuroscience. Psychol Bull 126:890–906
Dörner D (1999) Bauplan für eine Seele. Rowohlt, Hamburg
Dörner D, Bartl C, Detje F (2002) Die Mechanik des Seelenwagens. Eine neuronale Theorie der Handlungsregulation. Huber, Bern
Doya K (2008) Modulators of decision making. Nature Neuroscience 11(4):410–416
Dreisbach G, Goschke T (2004) How positive affect modulates cognitive control: reduced perseveration at the cost of increased distractibility. J Exp Psychol Learn Mem Cogn 30:343–353
Dreisbach G, Müller J, Goschke T, Strobel A, Schulze K, Lesch KP, Brocke B (2005) Dopamine and cognitive control: the influence of spontaneous eye-blink rate and dopamine gene polymorphisms on perseveration and distractibility. Behav Neurosci 119:483–490
Duncker K (1945) On problem solving. Psychol Monogr 58:5 Whole No. 270
Easterbrook JA (1959) The effect of emotion on cue utilization and the organization of behavior. Psychol Rev 66:187–201
Eriksen BA, Eriksen CW (1974) Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept Psychophys 16:143–149
Erk S, Walter H (2000) Denken mit Gefühl. Der Beitrag von funktioneller Bildgebung und Simulationsexperimenten zur Emotionspsychologie. Nervenheilkunde 1:3–13
Federmeier KD, Kutas M (1999) A rose by any other name: long-term memory structure and sentence processing. J Memory Lang 41:469–495
Federmeier KD, Kirson DA, Moreno EM, Kutas M (2001) Effects of transient, mild mood states on semantic memory organization and use: an event-related potential investigation in humans. Neurosci Lett 305:149–152
Fiedler K (2001) Affective states trigger processes of assimilation and accommodation. In: Martin LL, Clore GL (eds) Theories of mood and cognition: a user´s guide. Erlbaum, Mahwah, NJ, pp 85–98
Forgas JP (ed) (2000) Feeling and thinking: the role of affect in social cognition. Cambridge University Press, New York, NY
Fredrickson BL (2001) The of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. Am Sci 56:218–226
Fredrickson BL, Branigan C (2005) Positive emotions broaden the scope of attention and thought-action repertoires. Cogn Emot 19:313–332
Fredrickson BL, Joiner T (2002) Positive emotions trigger upward spirals toward emotional well-being. Psychol Sci 13:172–175
Gasper K (2003) When necessity is the mother of invention: mood and problem solving. J Exp Soc Psychol 39:248–262

Gasper K, Clore GL (2002) Attending to the big picture: mood and global vs. local processing of visual information. Psychol Sci 13:34–40
Goschke T (1996) Gedächtnis und Emotion: Affektive Bedingungen des Einprägens, Behaltens und Vergessens. In: Albert D, Stapf K-H (eds) Enzyklopädie der Psychologie. Serie II, Band 4: Gedächtnis. Hogrefe, Göttingen, pp 605–694
Goschke T (2000) Involuntary persistence and intentional reconfiguration in task-set switching. In: Monsell S, Driver J (eds) Attention and performance XVIII: control of cognitive processes. MIT Press, Cambridge, MA, pp 331–356
Goschke T (2003) Voluntary action and cognitive control from a cognitive neuroscience perspective. In: Maasen S, Prinz W, Roth G (eds) Voluntary action: brains, minds, and sociality. Oxford University Press, Oxford, pp 49–85
Goschke T (2007) Volition und kognitive Kontrolle. In: Müsseler J (ed) Allgemeine Psychologie (2. Auflage). Spektrum Akademischer Verlag, Heidelberg, pp 232–293
Gray JR (2001) Emotional modulation of cognitive control: approach-withdrawal states double-dissociate spatial from verbal two-back task performance. J Exp Psychol Gen 130:436–452
Isen AM (1999) Positive affect. In: Dalgleish T, Power MJ (eds) Handbook of cognition and emotion. Wiley, Chichester, pp 521–539
Isen AM (2004) Some perspectives on positive feelings and emotions: positive affect facilitates thinking and problem solving. In: Manstead ASR, Frijda N, Fischer A (eds) Feelings and emotions: the Amsterdam symposium. Cambridge, NY, pp 263–281
Isen AM, Daubman KA (1984) The influence of affect on categorization. J Pers Soc Psychol 47:1206–1217
Isen AM, Daubman KA, Nowicki GP (1987) Positive affect facilitates creative problem solving. J Pers Soc Psychol 52:1122–1131
Isen AM, Johnson MMS, Mertz E, Robinson GF (1985) The influence of positive affect on the unusualness of word associations. J Pers Soc Psychol 48:1413–1426
Kuhl J (1983) Emotion, Kognition und Motivation: II. Die funktionale Bedeutung der Emotionen für das problemlösende Denken und für das konkrete Handeln [Emotion, cognition, and motivation: II. The functional role of emotions in problem-solving and action control]. Sprache und Kognition 4:228–253
Kuhl J (2000) A functional-design approach to motivation and self-regulation: the dynamics of personality systems interactions. In: Boekaerts M, Pintrich PR, Zeidner M (eds) Handbook of self-regulation. Academic Press, San Diego, pp 111–169
Kuhl J (2001) Motivation und Persönlichkeit. Hogrefe, Göttingen
Kuhl J, Kazén M (1999) Volitional facilitation of difficult intentions: joint activation of intention memory and positive affect removes Stroop interference. J Exp Psychol Gen 128:382–399
Lang PJ, Bradley MM, Cuthbert BN (1998) International Affective Picture System (IAPS): Technical Manual and Affective Ratings (Univ. of Florida Center for Research in Psychophysiology, Gainesville, FL)
LeDoux JE (1996) The emotional brain. Simon and Schuster, New York
Luchins AS (1942) Mechanization in problem solving: the effect of Einstellung. Psychol Monogr 54:1–95
Martin LL, Clore GL (eds) (2001) Mood and social cognition: contrasting theories. Lawrence Erlbaum Associates, Mahwah, NJ
Mednick SA, Mednick MT (1967) Examiner's manual: remote associates test. Houghton Mifflin, Boston
Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202
Müller J, Dreisbach G, Brocke B, Lesch K-P, Strobel A, Goschke T (2007) Dopamine and cognitive control: the influence of spontaneous eyeblink rate, DRD4 exon III polymorphism and gender on flexibility in set-shifting. Brain Res 1131:155–162
Oaksford M, Morris F, Grainger B, Williams J, Mark G (1996) Mood, reasoning, and central executive processes. J Exp Psychol Learn Mem Cogn 22:476–492
Panksepp J (1982) Toward a general psychobiological theory of emotions. Behav Brain Sci 5:407–422

Phillips LH, Bull R, Adams E, Fraser L (2002) Positive mood and executive functions: evidence from Stroop and fluency tasks. Emotion 2:12–22

Piaget J (1975) Die Entwicklung des Objektbegriffs. In: Piaget J (ed) Der Aufbau der Wirklichkeit. Klett, Stuttgart, pp 14–99

Roberts AC (2008) Dopaminergic and serotonergic modulation of two distinct forms of flexible cognitive control: attentional set-shifting and reversal learning. In: Bunge SA, Wallis JD (eds) Neuroscience of rule-guided behaviour. Oxford University Press, Oxford, pp 283–312

Rowe G, Hirsh JB, Anderson AK (2007) Positive affect increases the breath of attentional selection. PNAS 104:383–388

Schwarz H (2001) Feelings as information: implications for affective influences on information processing. In: Martin LL, Clore GL (eds) Mood and social cognition: contrasting theories. Lawrence Erlbaum Associates, Mahwah, NJ, pp 159–176

Schwarz N (1990) Feelings as information: informational and motivational functions of affective states. In: Higgins ET, Sorrentino R (eds) Handbook of motivation and cognition: foundations of social behaviour, vol 2. Guilford Press, New York, pp 527–561

Schwarz N, Bless B (1991) Happy and mindless, but sad and smart? The impact of affective states on analytic reasoning. In: Forgas J (ed) Emotion and social judgment. Pergamon, Oxford, England, pp 55–71

Schwarz N, Clore GL (1983) Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. J Pers Soc Psychol 45:513–523

Schwarz N, Bless B, Bohner G (1991) Mood and persuation: affective states influence the processing of persuasive communications. Adv Exp Soc Psychol 45:161–199

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA.: MIT Press

Cultural Differences in Thinking Styles

Shihui Han

Abstract Recent cross-cultural psychological research showed ample evidence that humans from different cultures are characterized by divergent cognitive processing styles. Specifically, people from Western cultures (e.g., European Americans) are characterized by an analytic cognitive style that is attuned to salient focal objects but less sensitive to contexts, whereas people engaged in East Asian cultures (e.g., Chinese, Japanese, Korean) possess a holistic cognitive style that is attuned to background and contextual information. Brain imaging research showed that cultures not only shape multiple-level cognitive processes but also induce variation of neural correlates underlying cognitive processes such as perceptual/attentional processing. The findings help to understand how cognitive processes and the underlying neural mechanisms are modulated by cultures so as to give rise to cultural specific thinking styles.

1 Introduction

The world consists of cultures with tremendous differences that produce unique sociocultural contexts. People from divergent cultures not only behave in very different ways but think with different cognitive styles as well. Classical cognitive psychological research examines cognitive mechanisms underlying human mental processes without considering cultural influence and assumes that cognitive mechanisms uncovered in one cultural group can be applied to other cultures. However, recent cross-cultural psychological research has documented ample evidence for cultural differences in multiple-level human cognitive processes from perception to social cognition (Nisbett 2003). More recently, transcultural brain imaging studies further revealed cultural differences in neural mechanisms underlying multiple-level human cognition (Han and Northoff 2008) and thus provide a neural account of

S. Han
Department of Psychology, Peking University, 5 Yiheyuan Road, Beijing, 100871, China
e-mail: shan@pku.edu.cn

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_19,

cultural specific cognitive or thinking styles. In this chapter I first review social psychological evidence for cultural differences in perceptual attentional processing. I then review recent brain imaging studies that provide preliminary evidence for modulation of the underlying neural mechanisms by cultural differences. I finally discuss how these findings help us to understand the interplay between sociocultural contexts and the neural correlates of human cognition.

2 Cultural Differences in Perceptual and Attentional Processing

There is no doubt that sensory input determines to a large extent perceptual processing in the brain. The classical neurophysiological research investigated receptive field properties of neurons in the visual system by examining how neuronal responses are tuned by visual stimuli with specific features (e.g., orientation and color, Hubel and Wiesel 1962; Livingstone and Hubel 1984). Brain imaging studies also explore how neural responses in specific brain areas are modulated by sensory or perceptual features of stimuli such as motion (Tootell et al. 1995) or faces (Puce et al. 1995). However, human beings live in different natural environments and different sociocultural contexts. Cultures, as ongoing collective social processes that generate social, psychological, linguistic, material, and other resources, influence human development across life span (Li 2003). To what degree are neural substrates of human cognition influenced by sociocultural contexts?

This issue has not captured neuroscientists' attention until social psychologists found robust evidence that cultures influence human cognitive processes including low-level perceptual processing. The social psychological research was guided by the hypothesis that people from Western cultures (e.g., European Americans) are characterized by an analytic cognitive style that is attuned to salient focal objects but less sensitive to contexts whereas people engaged in East Asian cultures (e.g., Chinese, Japanese, Korean) possess a holistic cognitive style that is attuned to background and contextual information (Nisbett et al. 2001; Nisbett 2003). The initial cross-cultural research assessed whether the effect of contextual information on perceptual analysis of a target is different across Americans and Chinese using a rod-and-frame task (Ji et al. 2000). Subjects were presented with a rotating rod centered at a tilted frame and were asked to rotate the rod so that it oriented orthogonal to the earth's surface. Even though subjects were asked to ignore the frame, their performances were influenced by the titled frame. Most important, the authors found that Americans were more accurate than Chinese in aligning the rod. In other words, Chinese performances were affected by the contextual information to a larger degree relative to American performances. This is possibly due to that Chinese automatically paid more attention to the contexts than Americans and resulted in stronger contextual interference on the perceived orientation of the rod.

A similar cultural difference in perceptual processing was demonstrated between Americans and Japanese using a framed-line test (Kitayama et al. 2003).

This test was designed to examine both the ability to incorporate and the ability to ignore contextual information. Subjects were first shown a vertical line printed downward inside a square frame. They were then presented with another square frame that was larger, smaller, or of equal size relative to the first frame. In an absolute task, which required ignoring the contextual information, subjects were instructed to draw a line in the second frame with the same absolute length of the line in the first frame. In a relative task, which required incorporation of contextual information, subjects had to draw a line in the second frame so that the proportion of the line to the size of the frame was identical to the previous stimulus. If the analytic-holistic difference in cognitive style exists between American and Japanese, Americans should make less errors in the absolute task but more errors in the relative task and Japanese should show a reverse pattern. This was indeed observed by Kitayama et al. (2003), suggesting that Americans were better in ignoring contextual information whereas Japanese were better in incorporating contextual information with the target.

Other paradigms were also developed to reveal cultural differences in perceptual analysis of complex visual scenes. For example, Nisbett and Masuda (2003) showed Americans and Japanese two successive pictures of complex scenes. The pictures exhibited scenes to mimic either the object-salience of a Western city (or farm) or the field salience of an East Asian city (or farm). The two versions of the same pictures were presented rapidly and the second picture differed from the first one in either salient foreground objects or the relationships between objects and less salient background objects. Sensitivity to changes was measured by asking the participants to report the changes they had seen. The authors found that Americans performed better in detecting changes in salient objects whereas Japanese were better in finding changes in contexts. Similar results were also observed when using simple shapes (Masuda and Nisbett 2006).

Why does cultural difference in perceptual/attentional processing occur? A possible account is that perceptual environments are different in Western and East Asian societies and the interplay between environments and mind leads to cultural specific patterns of percpetual and attentional processing. To test this assumption, Miyamoto et al. (2006) randomly sampled pictures of scenes from cities in Japan and the United States. They then asked American and East Asian students to rate, in each picture, how ambiguous the boundary of each object is and how the scene is chaotic. It turned out that Japanese scenes were judged to be more ambiguous and complex than American scenes. A similar conclusion was obtained based on ratings using a computer program. Thus both subjective and objective ratings suggest that Japanese scenes are more ambiguous and complex than American scenes and may thus encourage perception of the context. The authors further primed Japanese and Americans using these pictures before subjects performed a change-blindness task with different pictures. Interestingly, for both cultural groups, exposure to scenes from Japanese cities resulted in better performance of detection of contextual changes relative to exposure to scenes from American cities. The results lend support to the proposal that culturally characteristic environments may afford cultural specific patterns of perception.

The difference in cognitive styles between Western and East Asian cultures goes beyond the perceptual/attentional processing. For example, prior to the observation of cultural difference in perceptual/attentional processing, social psychologists had shown evidence that the self-concepts or styles of self contrual differ greatly between Western and East Asian cultures (Markus and Kitayama 1991, 2003). Specifically, Western independent-self views the self as an autonomous entity separated from others and stresses individual internal attributes and thoughts in behavior. In contrast, East Asian interdependent-self emphasizes fundamental connectedness between individuals in a society and contingencies between the individual's behaviors and thoughts and actions of others. It appears that cultural differences in both perceptual/attentional processing and self-concept are consistent with the analytic/holistic dichotomy in thinking styles between Western and East Asian cultures (Nisbett et al. 2001; Nisbett 2003). However, the relation between cultural differences in multiple-level cognitive processes is unclear. One possibility is that the self, as a cognitive structure with executive functions to organize information processing (Kühnen et al. 2001), influences or determines the styles of cognitive processes. To assess this hypothesis, Kühnen and Oyserman (2002) examined whether processing can be switched toward context-independent or context-dependent styles by a self-construal priming (Gardner et al. 1999) that leads to more independent or interdependent self-construals. After the self-construal priming procedure that asked subjects to circle the independent (e.g., I, mine) or interdependent (e.g., we, ours) pronouns in an essay, subjects were presented with a Navon-type compound stimulus (Fig. 1, Navon 1977) and were asked to identify the local or global letters in the compound stimuli. Interestingly, subjects with independent self-construal priming responded faster to the local than to the global

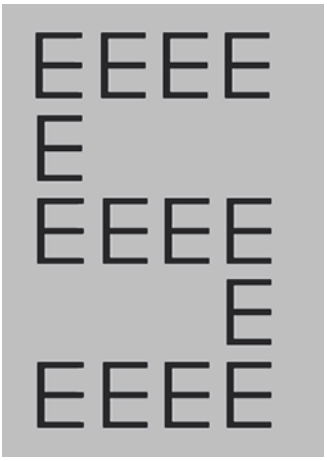

Fig. 1 Illustration of a Navon-type compound stimulus. A global S is made of local Es. Participants were asked to discriminate either the global and local letters

letters whereas a reverse pattern of performance was observed in subjects who were primed with interdependent self-construals. Kühnen et al. (2001) argued that self-construal priming results in a shift of processing mode with independent-self promoting a context-independent cognitive processing style and interdependent-self promoting a context-dependent cognitive processing style. Lin and Han (2009) further tested the cause-effect relation between the self-construals and the cognitive styles using a within-subjects design and a flanker task. The same group of Chinese subjects were first exposed to self-construals that primed the Eastern interdependent self or the Western independent self. They were then asked to discriminate a central target letter flanked by compatible or incompatible stimuli. Lin and Han found that, while responses were slower to the incompatible than compatible stimuli, this flanker compatibility effect was increased by the interdependent relative to the independent self-construal priming, suggesting that switching toward the interdependent self in mono-cultural participants results in increased scope of visual attention. This was further supported by a second experiment in Lin and Han's study (2009), which showed that the same group of subjects responded faster to the global than local targets in Navon-type compound letters after the interdependent self-construal priming whereas a reverse pattern was evident after the independent self-construal priming (Fig. 2).

Taken together, the aforementioned psychological studies indicate that different cognitive styles characterize Western and East Asian cultures. Specifically, the cognitive processes of Westerners are dominated by a context-independent style, paying more attention to salient focal objects and ignoring context information, whereas the cognitive processes of East Asians are characterized by a context-dependent style, paying more attention to background and contextual information.

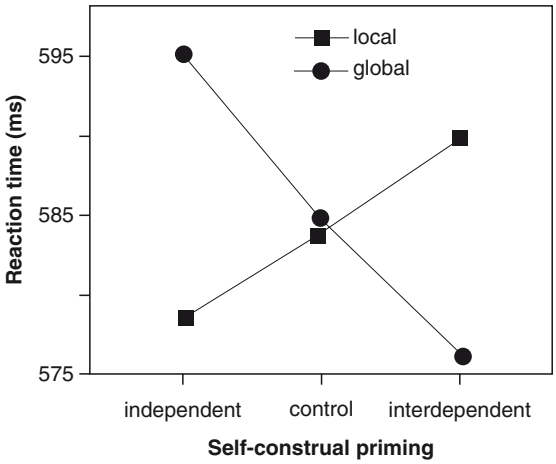

Fig. 2 Illustration of the modulation of responses to the global and local targets in Navon-type compound letters by self-construal priming in Lin and Han's (2009) study

Such cultural differences in cognitive styles are identified both in high- and low-level cognitive functions.

3 Neural Basis of Cultural Specific Cognitive Styles

While the human brain creates varieties of cultures, the brain itself is also shaped by external physical and sociocultural contexts (Han and Northoff 2008; Wexler 2006). Given the fact that human cognitive processes are influenced by cultures, an important question for neuroscientists concerns the intrinsic neural basis of cultural influence on human cognitive styles. Transcultural brain imaging provides an ideal tool to explore this issue.

Regarding cultural specific perceptual/attentional processing that was revealed by behavioral studies (e.g., Kitayama et al. 2003; Kühnen and Oyserman 2002; Lin and Han 2009), the modulation of underlying neural substrates may take place in different loci in the brain. For instance, the context-dependent cognitive style may be mediated by modulation of the activity in both the visual cortex engaged in perceptual processing and the neural circuit to guide attention. Indeed, recent transcultural brain imaging studies found evidence for both. In a recent functional magnetic resonance imaging (fMRI) study, Hedden et al. (2008) employed stimuli similar to those used by Kitayama et al. (2003). They presented Americans and East Asians with a series of stimuli that consisted of a vertical line inside a box. Participants were asked to perform either a relative-judgment task (judging whether the box and line combination of each stimulus matched the proportional scaling of the preceding combination) or an absolute-judgment task (judging whether the current line matched the previous line, regardless of the size of the accompanying box). As previous behavioral study suggested that Americans prefer the absolute-judgment task and East Asians prefer the relative-judgment task (Kitayama et al. 2003), Hedden et al. examined if sociocultural contexts and cultural practice result in decreased activity in the attentional network consisting of the frontal and parietal cortex in each cultural group when performing their preferred relative to unpreferred tasks. They first contrasted the unpreferred task in a difficult (both absolute and relative rules led to the same matching response) and an easy (the two rules led to opposing matching responses) condition and identified enhanced prefrontal and parietal activity in the difficult task. More interestingly, they found that, relative to East Asians, Americans showed greater activation in the parietal-frontal network when performing the relative-judgment task but weaker activation when performing the absolute-judgment task. However, the authors failed to find modulation of the neural activity in the early visual cortex. It seems that cultural influence led to modulation of the neural activity involved in attentional control rather than perceptual processing.

However, another fMRI study suggests that cultural influence may also generate modulation of the neural systems associated with focal object processing at an early stage of scene encoding. Goh et al. (2007) scanned both Americans and East Asians when the participants were shown picture stimuli of objects placed within background scenes. Successive picture stimuli could change in objects, background scenes, or both. The authors found that, in both cultural groups, repeated presentation of objects gave rise to reduced neural activity in lateral occipital cortex and repeated presentation of background scenes led to decreased activity in the parahippocampal gyrus. Nevertheless, such adaptation of neural responses was different between two cultural groups, i.e., Americans showed stronger adaptation responses in the lateral occipital areas than East Asians. Therefore, it was concluded that cultural specific experiences of object-focused visual processing in Americans result in stronger modulation of perceptual processes in the posterior visual cortex. Together with the findings of Hedden et al. (2008), these brain imaging findings indicate that cultural experiences not only adjust psychological processes and lead to cultural specific cognitive styles but also induce cultural specific patterns of neural processes. Apparently, cultures shape both the mind and the brain.

Would it be possible that cultural specific self-concepts are involved in modulation of neural substrates underlying cultural specific perceptual/attentional processing? It is difficult to test this by comparing brain imaging results from two cultural groups since aside from self-concepts there are also other differences between any two cultural groups. We recently investigated this using self-construal priming (Lin et al. 2008). Because prior research has shown evidence that cognitive styles are modulated by priming procedures that bias the self-concept toward independent or interdependent styles (Kühnen and Oyserman 2002; Lin and Han 2009), we tested if self-construal priming (Gardner et al. 1999) could modulate neural substrates underling the processing of global or local features of Navon-type stimuli. We first asked Chinese participants to read essays containing independent pronouns "I" or interdependent pronoun "We." Event-related brain potentials (ERPs) were then recorded when participants discriminated global or local letters of compound stimuli. We were particularly interested in whether early ERP components that arise from the visual cortex are modulated by independent and interdependent self-construal primes. Discrimination of both global and local letters elicited a positive activity peaked at about 100 ms after sensory stimulation with maximum amplitudes over bilateral visual cortex (P1), which has neural sources in the extrastriate visual cortex (Heinze et al. 1994). Most important, we found that the P1 amplitude was larger to local than global targets after the independent self-construal priming and a reverse pattern was observed following interdependent self-construal priming (Fig. 3). Our findings provide evidence that shift of cultural specific self-concepts led to changes of visual perceptual processing in extrastriate cortex. Thus it can be concluded that self-styles play an important role in generation of cultural specific cognitive styles.

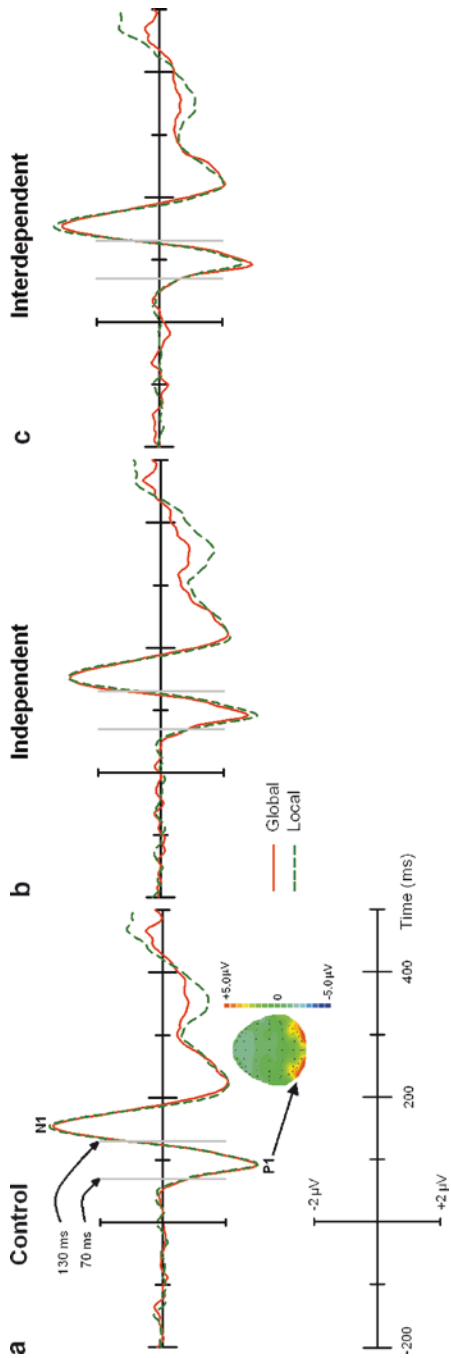

Fig. 3 Illustration of ERP results in Lin et al.'s (2008) work. ERPs elicited by global and local targets are shown respectively in (**a**) neutral, (**b**) independent, and (**c**) interdependent self-construal priming conditions. Voltage topographies illustrated the scalp distribution of the P1 component to global targets in the control priming condition

4 Conclusion

The findings of current cross-cultural psychological and transcultural brain imaging studies consistently support the assumption (or hypothesis) that Western cultures give rise to an analytic cognitive style whereas East Asian cultures result in a holistic cognitive style. The two different cultures shape both the cognitive processes and the underlying neural mechanisms. Although the findings reviewed in this chapter focus on perceptual/attentional processing, cultural differences in cognitive styles have been observed in other cognitive processes such as memory, language and music, causal attribution, mental state understanding, and self-awareness and self-representation. Cultural influence on multiple-level cognitive functions can result in cultural specific thinking styles, e.g., the analytic way of thinking in Western cultures versus the holistic way of thinking in East Asian cultures, which may then lead to cultural differences in social behaviors.

The emergence of cultural specific thinking styles reflects per se the interplay between ontogenesis and sociocultural contexts. The development of each individual or each brain is constrained by a sociocultural frame. Cultural specific cognitive styles and cultural specific neural underpinnings develop to adapt to a specific sociocultural context. However, the aforementioned findings do not exclude that different thinking styles can be observed in individuals in the same cultural group. Instead, the studies using cultural priming showed evidence that the cognitive or thinking styles of each individual can be biased toward different cultural patterns. The economic globalization has eased transfer and exchange among different cultures around the world. For instance, Chinese, particularly young Chinese students, are exposed to Western culture more frequently than ever before. East Asian cultures are also transferred to Western societies through media and immigrants. Consequently, each individual, to a certain degree, may possess multiple cultural knowledge though one culture dominates in a person. This provides the cultural basis of featuring multiple cognitive styles, which can be shifted by cultural priming.

The findings of cross-cultural psychological and transcultural brain imaging studies do not conflict with the existence of culture-invarant universal cognitive processes and neural correlates. For instance, Hedden et al. (2008) show that in judgment tasks the neural circuit consisting of frontal and parietal cortex is involved in both cultural groups . Only the magnitude of the neural activity varied between Western and East Asian cultures. Finally, it should be acknowledged that culture is a complex entity. Cultural differences are manifested in language, environment, social context, and subjective knowledge and belief. Future research may specify which aspect of cultures predominantly shapes cognitive styles and underlying neural mechanisms.

Acknowledgments This work was supported by National Natural Science Foundation of China (Project 30630025).

References

Gardner WL, Gabriel S, Lee AY (1999) "I" value freedom, but "we" value relationships: self-construal priming mirrors cultural differences in judgment. Psychol Sci 10:321–326

Goh JO, Chee MW, Tan JC, Venkatraman V, Herbrank A et al (2007) Age and culture moduldate object processing and object-scene binding in the ventral visual area. Cogn Affect Behav Neurosci 7:44–52

Han S, Northoff G (2008) Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. Nat Rev Neurosci 9:646–654

Hedden T, Ketay S, Aron A, Markus HR, Gabrieli DE (2008) Cultural influences on neural substrates of attentional control. Psychol Sci 19:12–17

Heinze HJ, Mangun GR, Burchert W, Hinrichs H, Scholz M et al (1994) Combined spatial and temporal imaging of brain activity during visual selective attention in humans. Nature 372:543–546

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154

Ji L, Peng K, Nisbett RE (2000) Culture, control, and perception of relationships in the environment. J Pers Soc Psychol 78:943–955

Kitayama S, Duffy S, Kawamura T, Larsen JT (2003) Perceiving an object and its context in different cultures: a cultural look at new look. Psychol Sci 14:201–206

Kühnen U, Oyserman D (2002) Thinking about the self influences thinking in general: cognitive consequences of salient self-concept. J Exp Soc Psychol 38:492–499

Kühnen U, Hannover B, Schubert B (2001) The semantic-procedural-interface model of the self: the role of self-knowledge for context-dependent versus context-independent models of thinking. J Pers Soc Psychol 80:397–409

Li SC (2003) Biocultural orchestration of developmental plasticity across levels: the interplay of biology and culture in shaping the mind and behavior across the life span. Psychol Rev 129:171–194

Lin Z, Han S (2009) Self-construal priming modulates the scope of visual attention. Q J Exp Psychol, Sec A 62(4):802–813

Lin Z, Lin Y, Han S (2008) Self-construal priming modulates visual activity underlying global/local perception. Biol Psychol 77:93–97

Livingstone MS, Hubel DH (1984) Anatomy and physiology of a color system in the primate visual cortex. J Neurosci 4:309–356

Markus HR, Kitayama S (1991) Culture and the self: implication for cognition, emotion and motivation. Psychol Rev 98:224–253

Markus HR, Kitayama S (2003) Culture, self, and the reality of the social. Psychol Inq 14:277–283

Masuda T, Nisbett RE (2006) Culture and change blindness. Cogn Sci 30:381–399

Miyamoto Y, Nisbett RE, Masuda T (2006) Culture and the physical environment: holistic versus analytic perceptual affordances. Psychol Sci 17:113–119

Navon D (1977) Forest before trees: the precedence of global features in visual perception. Cogn Psychol 9:353–383

Nisbett RE (2003) The geography of thought. Free Press, New York

Nisbett RE, Masuda T (2003) Culture and point of view. Proc Natl Acad Sci USA 100:11164–11170

Nisbett RE, Peng K, Choi I, Norenzayan A (2001) Culture and systems of thought: holistic vs. analytic cognition. Psychol Rev 108:291–310

Puce A, Allison T, Gore JC, McCarthy G (1995) Face-sensitive regions in human extrastriate cortex studied by functional MRI. J Neurophysiol 74:1192–1199

Tootell RBH, Reppas JB, Dale AM, Look RB et al (1995) Visual motion aftereffect in human cortical area MT revealed by functional magnetic resonance imaging. Nature 375:139–141

Wexler B (2006) Brain and Culture. MIT Press, Cambridge

Part V
Modeling and Neurobiological Aspects

Natural Selection in the Brain

Chrisantha Fernando and Eörs Szathmáry

Abstract This chapter explores the possibility that natural selection takes place in the brain. We review the theoretical and experimental evidence for selectionist and competitive dynamics within the brain. We propose that in order to explain human problem-solving, selectionist mechanisms demand extension to encompass the full Darwinian dynamic that arises from introducing replication of neuronal units of selection. The algorithmic advantages of replication of representations that occur in natural selection are not immediately obvious to the neuroscientist when compared with the kind of search normally proposed by instrumental learning models, i.e. stochastic hill-climbing. Indeed, the notion of replicator dynamics in the brain remains controversial and unproven. It starts from early thoughts on the evolution of ideas, and extends to behaviourist notions of selection of state–action pairs, memetics, synaptic replicators, and hexagonal cortical replicators. Related but distinct concepts include neural selectionism, and synfire chains. Our contribution here is to introduce three possible neuronal units of selection and show how they relate to each other. First, we introduce the Adams synapse that can replicate (by quantal budding) and mutate by attaching to nearby postsynaptic neurons rather than to the current postsynaptic neuron. More generally, we show that Oja's formulation of Hebbian learning is isomorphic to Eigen's replicator equations, meaning that Hebbian learning can be thought of as a special case of natural selection. Second, we introduce a synaptic group replicator, a pattern of synaptic connectivity that can be copied to

C. Fernando (✉)
MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London, UK;
Center for Computational Neuroscience and Robotics, Sussex University, Brighton, Falmer, UK;
Collegium Budapest (Institute for Advanced Study), Szentháromság u. 2, H-1014, Budapest, Hungary

E. Szathmáry
Collegium Budapest (Institute for Advanced Study), Szentháromság u. 2, H-1014, Budapest, Hungary;
The Parmenides Foundation, Kirchplatz 1, D-82049 Munich/Pullach, Germany;
Institute of Biology, Eötvös University, Pázmány Péter sétány 1/c, H-1117, Budapest, Hungary

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_20,

other neuronal groups. Third, we introduce an activity replicator that is a pattern of bistable neuronal activities that can be copied between vectors of neurons. This last type of replicator is not composed of the first two kinds, but may be dependent upon them. We suggest how these replicators may take part in diverse aspects of cognition such as causal inference, human problem solving, and memory.

Abbreviations

AMPA	α-Amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
L0	Layer 0 in the copying structure
L1	Layer 1 in the copying structure
MTT	Multiple trace theory of memory
NS	Natural selection
PCG	Polychronous group
PFC	Prefrontal cortex
RL	Reinforcement learning
STDP	Spike time dependent plasticity
TD	Temporal difference

1 Introduction

The neuronal replicator hypothesis proposes that plastic adaptive thought and behaviour arises from the replication and natural selection of units of selection within the brain. Units of selection are entities that multiply and exhibit hereditary variation. These units of selection are of three types: (1) synapses, (2) groups of synapses, and (3) patterns of bistable neuronal activity. We propose that groups of synapses (2) are relatively slow replicators responsible for the process of memory consolidation and are involved in the construction of internal causal models of the environment. Rapid dynamical neuronal (3) replicators may be responsible for the kind of search undertaken in working memory and creative thinking.

In this section, we review the history of ideas about natural selection in the brain. Section 2 proposes three plausible models of neuronal replicators. Section 3 discusses the superiority of natural selection above other non-deterministic search algorithms. Section 4 discusses the frame problem, i.e. how an initial population of replicators is to be initialized. Section 5 argues that fitness functions, i.e. measures of subjective utility, are themselves evolved within the brain. Section 6 demonstrates how search can be structured using Hebbian learning. Finally, Section 7 demonstrates the importance of structured search for generative search tasks, i.e. for tasks demanding creativity.

Variants of these ideas are so old that they are embedded in language. Dennett writes that “the Latin verb *cogito* is derived, as St. Augustine tells us, from Latin words meaning *to shake together,* while the verb *intelligo* means *to select among.*

The Romans, it seems, knew what they were talking about" (Dennett 1981). Along a similar theme, William James appeals to introspection in claiming that there is spontaneous variation of ideas in the brain most of which perish through their worthlessness (James 1890). James Baldwin again uses introspection to elaborate on the structured variability exhibited by thought; we do not generate thoughts randomly as "scatterbrains" do (Baldwin 1898). He writes "intelligence itself, in its procedure of tentative experimentation, or 'trial and error', appears to operate in accordance with it [Natural Selection]" (Baldwin 1909).

Anecdotal reports of the introspections of great thinkers contain claims of sudden insight based on similarity or identity, or of a process of collision and interlocking of pairs of ideas until they formed stable combinations, e.g. Poincaré (Hadamard 1945). The poet Paul Valery explicitly states that "it takes two to invent anything" referring according to Daniel Dennett to "a bifurcation in the individual inventor", with a generative part and a selective part (Dennett 1981). However, there is much post hoc fabrication in introspection.

Jacques Monod draws a parallel between the evolution of ideas and that of the biosphere: "Ideas have retained some of the properties of organisms. Like them, they tend to perpetuate their structure and to breed; they too can fuse, recombine, segregate their content; indeed they too can evolve, and in this evolution selection must surely play an important role. I shall not hazard a theory of the selection of ideas. But one may at least try to define some of the principal factors involved in it. This selection must necessarily operate at two levels: that of the mind itself and that of performance" (Monod 1971). Monod appeals both to introspection and to observation of idea dynamics in society.

Referring to creativity once again, Donald Campbell argues: "[A] blind-variation-and-selective-retention process is fundamental to all inductive achievements, to all genuine increases in knowledge, to all increases in the fit of system to environment". Furthermore, "in going beyond what is already known, one cannot but go blindly. If one can go wisely, this indicates already achieved wisdom of some general sort" (Campbell 1974).

Richard Dawkins proposes that "memes are neural circuits with phenotypic effects such as words, music, visual images, styles of clothes, facial or hand gestures, skills" (Dawkins 1982) that replicate between brains. The appeal is to more than introspection, it is to the supposed evidence for cultural inheritance and selection (Boyd and Richerson 2005). Aunger says that the existence of memes has not yet been proven (Aunger 2002). He extends Dawkins' hypothesis by proposing that memes were originally intra-brain replicators used by the brain for making "backups" of distributed functions, and for repair of functions following neuronal cell death. He suggests that for rapid replication they must be based on the *electrochemical states of neurons*, rather than on their connectivity. According to Aunger, a replicated neuromeme will result in the new copy having the "same potential to fire as the source ensemble after excitation". "A [neuro]meme does not depend on the identity of individual neurons; it also does not depend on the unique connectivity of some neurons to other neurons, nor its positional relationship to the cortex as a whole." Aunger proposes that a neuromeme existing in two parts of the brain

(e.g. visual cortex and auditory cortex) will receive a different set of afferents and produce a different set of efferents, so its function must be a general algorithm or general dynamical system that can "add value" independent of the local information-processing stream in which it finds itself being copied to. In other words, "the ability to slot into a wide variety of neuronal complexes, to be a good neural citizen, is a favoured quality" (p. 228). We can think of the neuromeme as an informational "chemical" capable of moving around the brain and reacting in many processes.

Jablonka and Lamb also propose the notion of replication and selection of ideas in the form of a symbolic inheritance system; however, they deny the existence of memes (Jablonka and Lamb 2005) because they claim that cultural units of replication/inheritance are not reproducers, i.e. they do not specify their own developmental machinery (Greisemer 2000). Certainly, the *Canterbury tales* is a unit of evolution in that it has multiplication, heredity and variation (Barbrook et al. 1998), but it does not have development. In other words, it does not self-referentially specify the means of its own production to the same extent that a gene specifies a polymerase, and thus it appears not to be capable of open-ended evolution (Bedau 1998). Here, we consider only the evidence for neuronal replicators *within* brains, not *between* brains.

Most of the above claims appeal to introspection and a call to the power of natural selection. In contrast, after a career of detailed analysis of human problem solving, Herbert Simon claims that "human problem solving … involves nothing more than varying mixtures of trial and error and selectivity" p. 97 (Simon 1996). The work of Simon and Newell was all about how humans solve problems *consciously* by devising goals and heuristics for search (Newell and Simon 1972). We share Marr's objection to their work. Boden writes that he saw "their apparently human-like programmes as superficial gimmickry, not explanation" p. 433, Vol I (Boden 2006) because he thought computational explanations referred to unconscious, automatic, modular processing rather than to conscious processes. We propose that to explain the unlimited productivity of thought and behaviour one must ultimately understand how combinatorial syntax and semantics, and the structure sensitivity of processes must be implemented in neuronal systems, rather than in a mind as "mental representations" (Fodor and Pylyshyn 1988; Harvey 2008).

Dennett claims that the plausibility of the kind of claims about the fundamental cognitive architecture come because they restate an abstract principle known as "the law of effect," which must hold for any possible adequate explanation of behaviour (Dennett 1981). At the maximum level of generality it states that behaviour tends to become adapted to its environment. Dennett points out that "the law of effect is closely analogous to the principle of natural selection" in that both principles aim to explain adaptation. Whereas natural selection deals with organisms, the Law of Effect deals with a population of stimulus–response pairs, where S–R pairs are selected that produce the highest expected long-term reward. However, he points to the inadequacy of various behaviourist instantiations of the Law of Effect in that they are single-level and not multi-level systems of natural selection. Since it was coined by Thorndike, the Law of Effect has appeared as Hull's "law of primary reinforcement" and Skinner's "principle of operant conditioning", and it has not been effective in explaining all of behaviour. In principle, it is the idea that a set of hard-wired reinforcers can select for a set of soft-wired "essentially arbitrary,

undesigned temporary interconnections" (Dennett 1981). The natural selection algorithm is implemented by what Dennett calls "Skinnerian creatures" only by "actual behavioural trial and error in the environment". It follows that if there can be an environmental emulator (a feedforward model of the environment that can return an expected reward value given a potential behaviour as input), it is possible to simulate reinforcement from the environment without actually having to interact with it behaviourally, and that natural selection at the organismal level could select for appropriate emulators. The emulator can be considered to be a kind of internal value system that implements selection in lieu of external reward objects, for instead of requiring the environment to provide reward it provides the rewards that would be expected from the environment given a certain action. The emulators themselves need not be hard-coded; instead, there may be a set of genetically specified values for selecting soft-coded emulators. The same principle can establish a hierarchy of intrinsic selection emulators. Dennett quotes Popper that indeed such a hierarchical set of emulators and selection mechanisms allows "hypotheses to die in our stead". Thus, Dennett provides an argument not from introspection, but from logical necessity that adaptive behaviour and cognition necessarily have the same class of explananda as organismal natural selection, i.e. adaptation. Cooper has similarly argued that a wide range of rational behaviour reduces to natural selection, and that the concept of fitness can in some cases be replaced by the concept of subjective utility (Cooper 2001).

Given that a class of stochastic search processes can explain the production of adaptation, what are the different sub-classes of these processes, what are the theoretical constraints on such processes, and how can these processes be implemented in the brain? One general constraint is the number of tests (e.g. utility assessments) that can be carried out in a given period of time. Very generally, one can say that given such a constraint, a system benefits if it can structure variability such that it generates the variants most likely to be adaptive. There are three other general constraints that determine the outcome and efficiency of a non-deterministic search: (1) selection criteria, i.e. how (unconscious) subjective utility is assigned to a solution; (2) the neuronal representation of solution space; and (3) how solutions are assigned neuronal search resources based on the current neuronal assessment of subjective utility. A full understanding of the search involves a full algorithmic description of each of these four aspects and how they are neuronally implemented. The above constraints come from experience with genetic algorithms and artificial selection for optimization purposes. We assume the brain to be a device that undertakes such a non-deterministic search, and we explore specifically the hypothesis that the class of non-deterministic search processes it implements is a kind of natural selection.

2 How Could Neuronal Natural Selection Work?

Units of evolution are entities that multiply, are capable of stably transmitting variation across generations, but are also subject to variability, i.e. offspring do not perfectly resemble their parents. If such entities have differential fitness, then natural selection

generates adaptation (Maynard Smith 1986; Muller 1966). Replication can establish covariance between a phenotypic trait and fitness, a fundamental requirement for adaptation by natural selection (Price 1970). The natural selection algorithm (Dennett 1995) can have many possible implementations (Marr 1983), for example units of evolution include: some units of life (Gánti 2003) such as organisms and lymphocytes evolving by somatic selection (Edelman 1994), but also informational entities (without metabolism) such as viruses, machine code programmes (Lenski et al. 1999) and binary strings in a genetic algorithm (Mitchell 1996). We propose three different implementations of neuronal natural selection at different spatial and temporal scales.

2.1 Synapses

There are two ways of thinking of synapses as replicators. The first is described by Paul Adams who claims that synapses replicate and are the minimal neuronal units of selection. Synapses replicate by increasing the amount of quantal (i.e. discrete) release from the pre- to the post-synaptic neuron, according to a Hebbian rule. That is, if the pre- and post-synaptic neurons fire together, a uni-quantal synapse can become a bi-quantal synapse. Mutations are noisy quantal Hebbian learning events where a synapse is made to contact an *adjacent* post-synaptic neuron rather than to enhance the connection to the current post-synaptic neuron (Adams 1998). Synapses compete with each other for correlation resources, and other reward resources if for example dopamine acts to modulate Hebbian learning. Adams demonstrates how arrays of synaptic weights can be selected, and how error-correction mechanisms in cortical layer VI can adjust the synaptic mutation rate.

The second way to think of synapses is as implementing a special case of Eigen's replicator equation. This is a continuous-time, continuous state-space, deterministic model. The original Eigen equation was applied to very large populations of molecules, so that molecular numbers could be replaced by continuous concentrations, despite the underlying particulate nature of macromolecules. By the same token, the Oja rule assumes that synaptic strength changes continuously, despite the underlying, necessarily particulate, nature of the molecules involved. We have demonstrated a mathematical isomorphism between Hebbian learning and Eigen's replicator equations, a standard *model* of natural selection dynamics in chemical and ecological systems. We briefly review the nature of this isomorphism. Hebbian learning can be said to *select* between synaptic weights on the basis of correlations in activity. Consider the Oja-version of Hebbian learning (Oja 1982) and a model of evolving template replicators (Eigen 1971). The Eigen equation is shown below:

$$\frac{dx_i}{dt} = A_i Q_i x_i + \sum_{j \neq i}^{N} m_{ij} x_j - \frac{x_i}{c} \sum_{j=1}^{N} \sum_{k=1}^{N} m_{ij} x_j, \tag{1}$$

where x_i is the concentration of sequence *i (*of RNA for example), m_{ij} is the mutation rate from sequence *j* to *i*, A_i is the gross replication rate of sequence *i* and Q_i is its copying fidelity, N is the total number of different sequences, and formally $m_{ii} = A_i Q_i$ (Eigen 1971). The negative term introduces the selection constraint, which keeps total concentration constant at the value of c (which can be taken as unity without loss of generality). The equation describes a set of templates in a reactor. The templates can replicate but can also mutate into each other.

The neurobiological model in question assumes rate coding (hence we are not dealing with spiking neurons) and multiplicative normalization (Oja 1982):

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} - \alpha v^2 \mathbf{w}, \tag{2}$$

where $\mathbf{w}$ is the synaptic weight vector, v is the output rate, $\mathbf{u}$ is the input rate vector, and the rest are constants. This rule enforces competition, and the square is apparently due to the fact that weights can also be negative. If the firing rate model is linear, and if we assume that the network is allowed to attain its steady state activity during training, and that the processes of synaptic plasticity (at the "the population genetic time scale") are slower than the dynamics of firing rates (at the "the ecological time scale"), then we have (Dayan and Abbott 2001).

$$v = \mathbf{w} \bullet \mathbf{u}, \tag{3}$$

where the sign · is the dot (or scalar) product, understood as:

$$\mathbf{w} \bullet \mathbf{u} = \sum_{b=1}^{N_u} w_b u_b \tag{4}$$

where N_u is the number of input synapses to the single neuron considered. Substituting (4) into (2) we obtain:

$$\tau_w \frac{d\mathbf{w}}{dt} = (\mathbf{w} \bullet \mathbf{u})\mathbf{u} - \alpha(\mathbf{w} \bullet \mathbf{u})^2 \mathbf{w}. \tag{5}$$

It is known that the quantity $|\mathbf{w}|^2 = \mathbf{w} \bullet \mathbf{w}$ relaxes over time to $1/\alpha$, hence the competition constraint. It is important that a synapse updates itself only based on locally available information (u_i, w_i and v). Note that if $\mathbf{w} \geq 0$ holds (this is the only real limitation of this study), then one can drop the square, and thus have

$$\tau_w \frac{d\mathbf{w}}{dt} = (\mathbf{w} \bullet \mathbf{u})\mathbf{u} - \alpha(\mathbf{w} \bullet \mathbf{u})\mathbf{w}. \tag{6}$$

This form actually ensures that non-negative weights remain non-negative. Now we write the above equation in a different form:

$$\tau_w \frac{dw_i}{dt} = u_i{}^2 w_i + u_i \sum_{j \neq i}^{N_u} u_j w_j - \alpha w_i \sum_{j=1}^{N_u} u_j w_j, \tag{7}$$

which looks almost like the Eigen equation with N_u different, mutationally coupled replicators. τ_w is just a scaling factor (the learning rate) so without loss of generality it can be taken as unity, and $1/\alpha$ can take the role of total concentration. Then u_i^2 is analogous to the fitness of unit *i*, and $u_i u_j$ is analogous to the mutation rate from sequence *j* to *i*. To make the analogy as neat as possible we divide all terms in (7) by $U^2 = \left(\sum_{j=1}^{N_u} u_j \right)^2$, and we introduce the new parameters $z_i = u_i/U$. By also taking $\tau_w = 1$ one can rewrite (7) as:

$$\frac{dw_i}{dt} = z_i{}^2 w_i + z_i \sum_{j \neq i}^{N_u} z_j w_j - \alpha' w_i \sum_{j=1}^{N_u} z_j \sum_{j=1}^{N_u} z_j w_j, \tag{8}$$

where it holds that $\sum_{j=1}^{N_u} z_j = 1$ and $\alpha' = \alpha/U$; *the latter,* just means that the total concentration is changed (without loss of generality). After rearrangement we obtain:

$$\frac{dw_i}{dt} = \sum_{j=i}^{N_u} z_i z_j w_j - \alpha' w_i \sum_{j=1}^{N_u} \sum_{k=1}^{N_u} z_j z_k w_k, \tag{9}$$

which is now *exactly* isomorphic to the Eigen equation (1). Here z_i^2 is the fitness of unit *i*, and $z_i z_j := m_{ij}$ is the mutation rate from sequence *j* to *i*. Note that, as for the dynamics of sequences in the simplest case, $m_{ij} = m_{ji}$. In terms of the Eigen equation it would hold that $z_i z_i = A_i Q_i$, so for Hebbian dynamics $A_i = Q_i$. Due to the definition of z_i the conservation relation (Eigen 1971) known for the Eigen equation also holds:

$$z_i(1 - z_i) = \sum_{j \neq i} z_i z_j. \tag{10}$$

Due to its internal structure (9) is a *special case* of the general Eigen equation. For example, in contrast to the molecular case, here the off-diagonal elements (the mutation rates) are not by orders of magnitude smaller than the diagonal ones (the fitnesses). Moreover, mutational coupling between two replicators is strictly the product of the individual fitnesses. Total concentration $c = U/\alpha$, i.e. it is proportional to the sum of individual fitnesses. This means, in contrast to the original Oja formulation (5) that a neuron under heavy fire allocates more resources to its synapses. For a fixed sum of input rates there is strict competition, and since rates cannot grow indefinitely for biophysical reasons, (9) cannot "blow up".

In short, the Eigen's equation can simulate Hebbian dynamics with the appropriate parameter values, but the reverse is not generally true: Oja's rule could not, for example, simulate the classical molecular quasispecies of Eigen in general. This hints at the more general possibility that although formal evolutionary dynamics could hold widely in brain dynamics, it is severely constrained in parameter space so that the outcome is behaviourally useful. Here, one can see an analogy to the immune system, which uses evolutionary dynamics in a specially constrained form.

A remarkable outcome of the above derivation is that although there was no consideration of "mutation" in the original setting, *there are large effective mutation rates* in (9)*:* this coupling ensures correlation detection between the units (synapses or molecules). Hence, if a molecular or biological population with dynamics (9) did exist, it would detect correlation between individual fitnesses. (Note that "synaptic mutation" *sensu* Adams (1998) refers to the erroneous formation of a new synapse on a target neuron different from where it should have landed as a result of strengthening; a process not considered here.) Neurons obeying the Oja rule can detect principal components of input vectors. In effect, what counts in the dynamics is $\langle v\mathbf{u} \rangle$, where $\langle \cdot \rangle$ denote averages over the ensemble of input patterns presented during training. Under the assumption of the above derivation the input correlation matrix $\mathbf{Q} = \langle \mathbf{uu} \rangle$ governs the dynamics, which is exactly the matrix that figures in the analogous Eigen equation (9) after scaling. Provided $\mathbf{u} > 0$ holds, the asymptotic behaviour (Thompson and McBride 1974; Jones et al. 1976) of the Eigen equation (the synaptic weight vector) finds the dominant eigenvector of the input correlation matrix.

In both Adams' and our formulation, the synapse is a unit of selection that detects and grows on the basis of *local* correlations. The whole idea becomes much more exciting when one considers possible "major transitions" (Maynard Smith and Szathmáry 1998) in neuronal evolutionary dynamics. These may be indispensable for complex thinking operations that operate on more than local correlations. We agree with Okasha that multiplication is crucial in this formulation (Okasha 2006). Next we discuss how selection could manifest itself at these higher levels of organization than synapses, namely, groups of synapses.

2.2 Groups of Synapses

We propose that an important higher-order unit of selection in the brain is a group of synapses of a particular pattern. Another way to think of this is as a pattern of neuronal connectivity. We have produced a model to show how a synaptic pattern could copy itself from one set of neurons to another. Before describing a biologically plausible mechanism for the replication of synaptic patterns, we clarify that this is distinct from the "copying" of receptive fields that has already been demonstrated (Song and Abbott 2001; Van Rossum et al. 2000; Young et al. 2007). William Calvin's proposal of hexagonal neuronal replicators is not discussed here

for it only does half the job of copying neuronal connectivity, i.e. it proposes a new means for copying receptive fields but does not explain reconstruction of connectivity (synaptic) patterns (Calvin 1987, 1996). By analogy to DNA, it establishes hydrogen bonds between strands, but does not explain how phosphodiester bonds within the new strand are formed.

Imagine that in a "parental" layer there exists a group of neurons connected by a particular synaptic pattern. Our aim is to copy this pattern into the neurons of another "offspring" layer. Our mechanism of connectivity copying depends on topographic map formation, i.e. projections that preserve local adjacency relationships between neurons within layer 1 in layer 2 (Song and Abbott 2001; Willshaw and von der Malsburg 1976) to establish receptive field correspondences between adjacent layers, coupled with spike-time dependent plasticity (STDP), an asymmetric kind of Hebbian weight change rule (Markram et al. 1997) to reconstruct connectivity patterns in the offspring layer. Neuronal resetting (Crick and Mitchison 1995) is also needed to erase connectivity so that a layer can be reused. The offspring layer copies the parental layer by implementing a causal inference algorithm neuronally (Pearl 2000). This algorithm uses STDP, but also requires additional topological error correction neurons to increase copying fidelity by comparing the firing of corresponding neurons in parent and offspring layer, and either increasing or decreasing afferents to the offspring neuron depending on whether it fires too often or too little. Also, an anisotrophic activity reverberation limitation mechanism is needed to increase copying fidelity by limiting non-local spread of activation in the offspring layer whilst still permitting plasticity and responsiveness to inputs from the parent layer, see Fig. 1 and its caption for a description of the mechanism. Recently, it has been found that acetylcholine can serve this role (Hasselmo 2006). Full details of the algorithm are available in Fernando et al. (2008). Note that all the component processes are individually well established in neuroscience.

The capacity of the system to copy neuronal topology is demonstrated in computer simulation using Izhikevich's model of spiking neurons (Izhikevich 2007). A fixed "parent" topology is initialized in layer 0 (L0). This is the topology we wish to copy to layer 1 (L1). We stimulate L0 randomly and sparsely (1–4 Hz). If a strong weight exists between two neurons in L0 there will be fixed cross-correlation (Dayan and Abbott 2001) between the firing of these neurons. Due to the assumption of a topographic map between L0 and L1, neurons in L1 will share a similar pattern of cross-correlation to neurons in L0. The weights between neurons in L1 are initially set to all be random and weak. This cross-correlation can be detected and used by the synapses in L1 to infer the underlying pattern of connectivity in L0. This is possible due to STDP (Song and Abbott 2001), which is an asymmetric synaptic weight change rule that increments the weight between two neurons if a presynaptic spike arrives prior to the firing of the postsynaptic neuron, and decrements the weight if a postsynaptic neuron fires before a presynaptic neuron.

We noted that systematic errors were made in copying because from cross-correlation alone, connectivity patterns can be Markov equivalent, i.e. two patterns of connectivity in L0 can produce the same cross-correlation pattern in L1, and STDP cannot distinguish between the two. Therefore, we introduced two error-correction

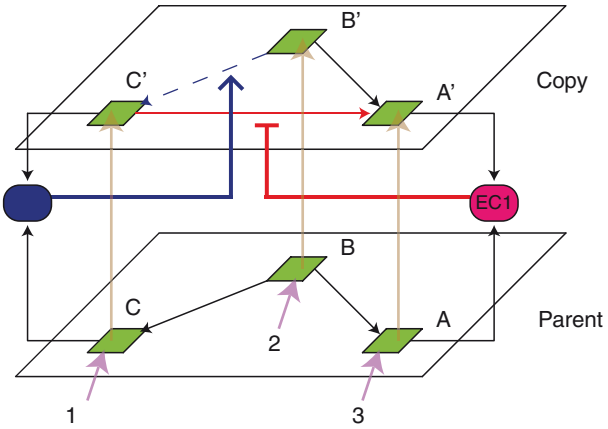

Fig. 1 An outline of the neuronal topology replication mechanism with error correction. The parental layer is on the *bottom*. The offspring layer is on the *top*. In this example, each layer has three neurons. Topographic collaterals (*vertical arrows*) connect parent to offspring layer. Copying is the reproduction of the intra-layer topology of the parent layer to the offspring layer. Error correction mechanisms are shown. STDP operates in the offspring layer. There are two error correction mechanisms; *EC1 (Right)* is a false positive error correction mechanism implemented using "observer" neurons (EC1) that negatively neuromodulate synapses onto neuron A' in the copy layer on the basis of differences in firing between the parental (A) and copy (A') layer neuron. We assume C is undergoing stimulation (1) when EC1 acts. *EC2 (Left)* is a false negative error correction mechanism (EC2) implemented using "observer" neurons that positively neuromodulate inputs that pass to a poorly firing neuron (C′) in the copy layer from the neuron that is undergoing stimulation (in this case we assume B is undergoing stimulation (2)) when EC2 acts. EC1 and EC2 type neurons are required for each neuron pair

mechanisms (Fig. 1) that compared the "phenotypes" of the two networks, i.e. the spike timings, and made directed changes the synaptic strengths between neurons. Error correction works by detecting neuronal spike-mispairing between layers and modifying afferent weights in the copy layer accordingly. These methods attempt to remove false positive connections and increase weights where there are false negative connections.

As Fig. 2a, b show the fidelity of copying is still not very high for some 3-neuron motifs, although it is almost perfect for others. In order to allow copying of larger motifs, activity reverberation within a layer has to be prevented. If this is not done, then STDP within the offspring layer makes causal inference errors because there are many possible pathways of activation that can account for a given cross-correlation pattern, and this number grows with the size and connectivity of the network. During the copy procedure, only the parental layer is stimulated with one spike at a time at frequency 1–4 Hz, a typical value for cortical neurons (Izhikevich et al. 2004). The source of depolarizing current to each neuron is classified as either intra-layer, I_i (from afferent neurons within the same layer), or inter-layer, I_e (from afferents outside the layer). If $I_i/I_e > \lambda$, where $\lambda = 0.1$, then the postsynaptic neuron does not send a spike to neurons within the same layer, but does send a spike to

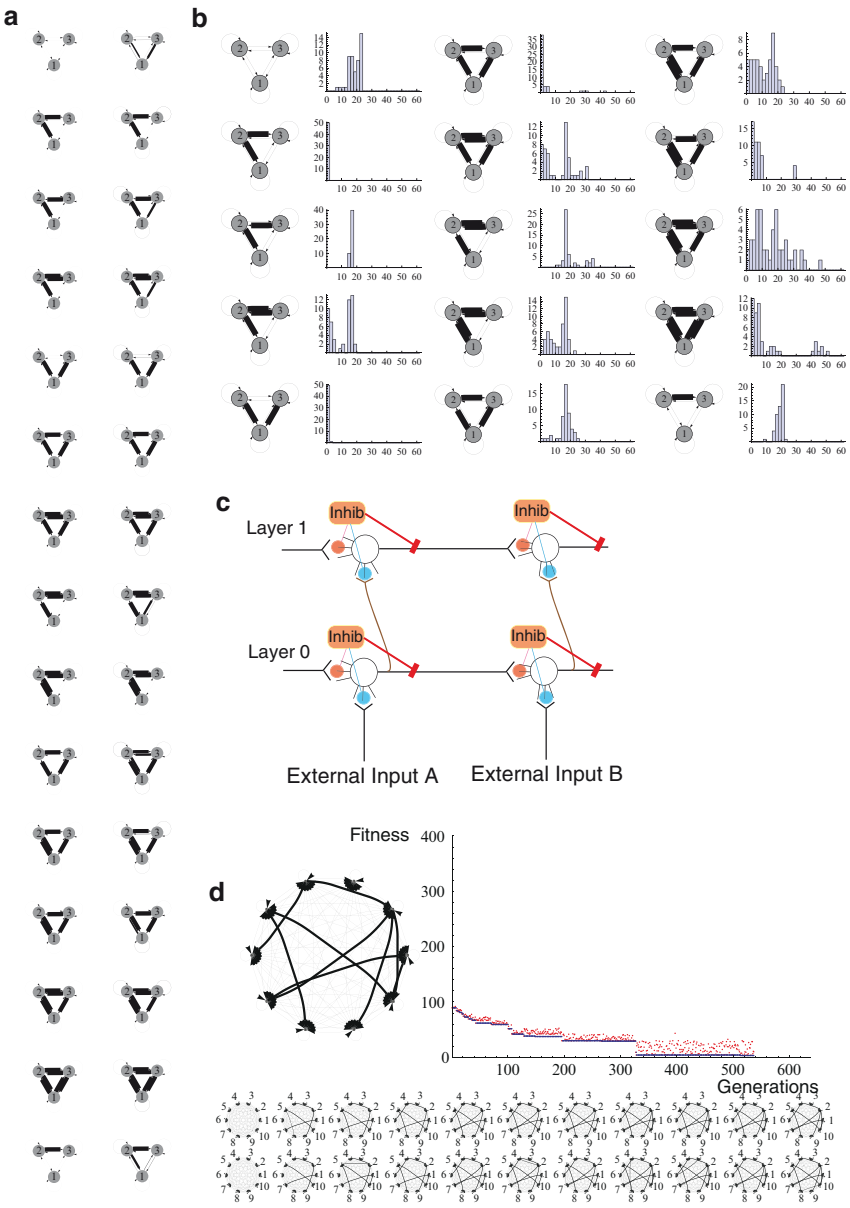

Fig. 2 **(a)** Parental 3-node motifs (*left*) and typical copies (*right*) for all 3-node motifs. (**b**) Distribution of Euclidean distances of copies from parent. 30 units = one maximum strength weight. Some motifs are copied much better than others. **(c)** Activity reverberation is implemented by inhibitory neurons associated with each excitatory neuron. This inhibitory neuron classifies depolarization of the associated excitatory neuron has being either from inside the layer or outside the layer. If most activation is from inside the layer, it blocks the spike from its associated excitatory neuron along the intra-layer efferent axon collateral. (**d**) Using activity reverberation, high-fidelity copying of a 20 node network can be undertaken, allowing a 1+1 ES to evolve a desired topology (*Top Left*). The fitness of the parent and offspring layers is shown (*Top Right*), along with the structure of samples of parent and offspring (*Bottom*) throughout the trial

neurons in other layers (i.e. passes the signal vertically but not horizontally). This ensures that if most of the current causing the neuron to fire is from a neuron within the same layer, the postsynaptic neuron does not pass this signal onto other neurons within the same layer, but only along vertical fibers to neurons in other layers. Despite this, we allow STDP to occur at all intra-layer synapses onto the postsynaptic neuron. On the other hand, if the current is mainly external, then a spike is produced that passes to both intra-layer and inter-layer neurons. The effect of this modification is to force only *local* horizontal spread of activation. A mechanism for achieving this is shown in Fig. 2c. This procedure eliminates causal inference errors that are made due to non-local correlations, and allows larger networks to be copied and evolved using, for example, a neuronally implemented 1 + 1 Evolution Strategy (Beyer 2001) to achieve a desired topology (see Fig. 2d).

The synaptic group replicator has a greater capacity for information integration than the synaptic replicators of which it is composed. It can not only grow on the basis of local synaptic correlations, but can respond specifically to temporal sequences of inputs, e.g. as polychromous groups (Izhikevich 2006).

2.3 *Patterns of Bistable Neuronal Activity*

We discuss two types of dynamical replicator, the first extends the concept of the synfire chain and the second depends on neuronal bistability. A synfire chain is a feed-forward network of neurons with several layers (or pools). Each neuron in one layer feeds many excitatory connections to neurons in the next pool, and each neuron in the receiving layer is excited by many neurons in the previous one. When activity in such a cascade of layers is arranged like a packet of spikes propagating synchronously from layer to layer it is called a synfire chain (Abeles 1982, 1991). There have been reports in the literature about observations of precisely repeating firing patterns (Abeles and Gat 2001). An excellent summary is provided by Abeles et al.[1] Figure 3 shows a synfire chain.

Figure 4 shows the temporal pattern of propagation of the spike packet:

Now consider the following arrangement (Fig. 5).

What we see is replication (with multiplication) of the spike packet. Note that since nothing is exactly accurate, undoubtedly there will be “mutations” with a certain frequency. If we have chains like this, then one can use it for the spread of spike packets according to rewarded information. If we imagine a lattice where every arrow between the neuronal groups can work both ways but in a reward-gated fashion then fitter and fitter packets can fill up the lattice. The snag, for the time

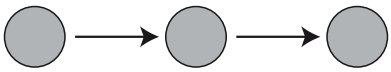

Fig. 3 A synfire chain in diagrammatic form. The *shading* indicates information propagated along the chain

[1] http://www.scholarpedia.org/article/Synfire_chain

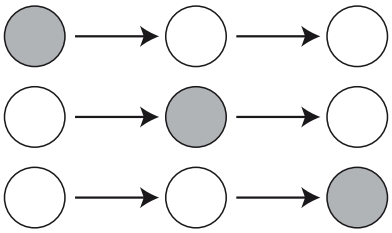

Fig. 4 Propagation of the spike packet along the synfire chain in time

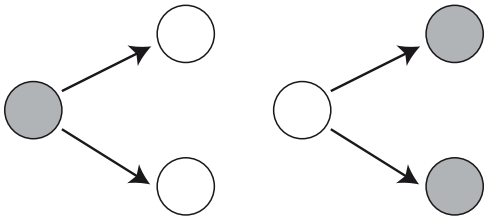

Fig. 5 A bifurcating synfire chain that can replicate patterns

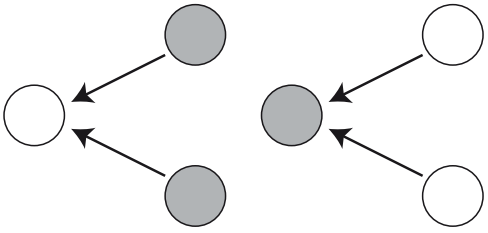

Fig. 6 Recombination of two patterns from two merging synfire chains

being, is the limited heredity potential due to the limited information a spike packet can propagate. Of course, recombination is easy to imagine, when two roughly equally fit packets want to be transmitted to the same neuron group (Fig. 6).

Stable propagation of the recombinant packets requires either many more stable spike packets than hitherto imagined, or that the incoming spike packets are by themselves *unstable.* Cateau and Fukai (2001) note that in the latter case the arrangement would work as an AND gate! The latter option highlights the potential importance of unstable patterns, as far as the chains are short enough so that the unstable spike patterns do not settle down to equilibrium. When two stable packets arrive at the same time recombination (propagation of a third-packet form) is possible. When the coincidence between two different spike patterns is not good, they are not stabilized. If the same two spike patterns arrive with good coincidence the latecomer is depressed; if the coincidence is low then they can propagate independently (Cateau and Fukai 2001).

These lessons are strongly suggestive for future work, if only unlimited heredity can be achieved.

Our second proposed dynamical replication mechanism is based on spiking neurons capable of bistability. The minimal unit of dynamical neuronal replication consists of a bidirectional coupled and gated pair of bistable neurons (Izhikevich 2007) (see Fig. 7).

Grouping these pairs together, one can make two layers coupled initially by a topographic map. Let us assume that the parental layer has neurons initialized randomly. If a bistable neuron is given some depolarizing current, it begins to fire repeatedly, whereas if a bistable neuron is given some hyperpolarizing current (in the correct phase) it stops firing. The state of neurons in the offspring layer is reset (i.e. all neurons are hyperpolarized to switch off spiking). Activity gates are opened for a brief period from the parental layer to the offspring layer, allowing the spiking neurons in the parental layer to switch on the corresponding neurons in the offspring layer. Activity gates between layers are then closed. Thus, the vector of activities in the parental layer is copied to the child layer. As in Fig. 2, a 1+1

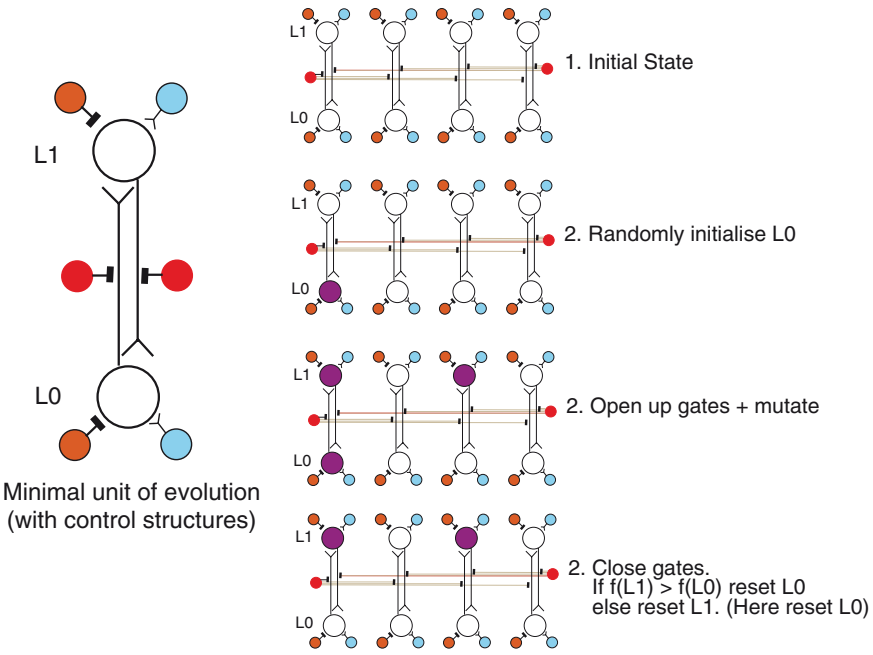

Fig. 7 An outline of how dynamical neuronal replicators can implement a 1+1 evolutionary strategy. The two bistable neurons are shown as *large circles* (*light* = not firing, *dark*= firing). They are coupled bidirectionally by gated axons. Each neuron can be reset and activated by external control. An array of such pairs is shown on the right. Initially, the vector of neurons is all off. Layer 0 is randomly initialized; here the neuron on the left becomes active. The up gates are opened allowing that neuron to activate its corresponding neuron in layer 1. The gates are closed and the fitness of each layer is assessed. The least fit layer is reset, and is overwritten by the fitter layer

Evolution Strategy can be implemented. That is, the two layers have their fitness assessed, the higher fitness layer is defined as the parent and the offspring layer activities are reset. The mechanism by which fitness is assessed may be via environmental interaction or by use of an emulator. The parent is then copied with mutation to the just reset offspring layer. The generation time may be less than a second in duration.

The bistable activity replicators are not composed of synapses or groups of synapses, but rather depend on synapses for their transmission. There is no satisfactory genetical analogy for this relationship between activity and connectivity replicators. Activity replicators should not be considered a higher-level unit, but rather a kind of rapid episynaptic unit, that depends on patterns of synaptic connectivity for their transmission. In that sense it is loosely analogous to epigenetic transmission of genes in OFF and ON states (cf. Jablonka and Lamb 2005).

3 Natural Selection in Comparison with Other (Neuronal) Stochastic Search Algorithms

It must be acknowledged that natural selection is only one of a set of stochastic search algorithms that could be implemented in the brain. This section examines other stochastic search algorithms, and how these compare with natural selection, for the purposes of explaining adaptive cognition and behaviour.

3.1 Neuronal Selectionism

Neuronal selectionism (or neuronal group selection) is often confused with natural selection because of Gerald Edelman's use of the term Neural Darwinism to describe the theory of neuronal group selection even though there is no multiplication of groups in his theory, and therefore no unit of evolution, and hence no natural selection (Crick 1989, 1990).

Changeux was the first to develop a neuronal selectionist theory (Changeux et al. 1973). Influenced by Donald Hebb's notion of a neuronal assembly (Hebb 1949), he describes a structural and functional unit of thought as the "mental object, a physical state, created by correlated, transient activity, both electrical and chemical, in a large population or "assembly" of neurons in several specific cortical areas" (p. 137, ibid). Anticipating the concept of "polychronous groups" (Izhikevich 2006), "a given neuron can take part in several graphs of different mental objects". Mental objects are described as interacting, with the wiring pattern of the cerebral machinery imposing a "grammar" on the linking of mental objects; the "grammar" of linking and combining concepts being very different from the "grammar" of primary percepts. Stabilization (learning) of a mental object occurs because the

"brain spontaneously generates crude, transient representations with graphs that *vary* from one instance to another. These pre-representations, exist *before* the interaction with the outside world. They arise from the recombination of preexisting sets of neurons or neuronal assemblies, and their diversity is thus great. On the other hand, they are labile and transient. Only a few of them are stored. This storage results from a *selection!* Darwin helps to reconcile Fodor with Epicurus! Selection follows from a *test of reality.* The test of reality consists of the comparison of a percept with a pre-representation. The test may involve resonance or, on the contrary, dissonance, between the two neuronal assemblies" p. 139 (Changeux 1985). No explanation is given as to how a beneficial property of one group would be transmitted when it is "recombined" with another group. The reticular formation is proposed to be responsible for the selection, by reentry of signals from cortex to thalamus and back to cortex, which is a means of establishing resonance between stored mental objects and percepts.

Changeux assumes the formation of pre-representations occurs spontaneously from a large number of neurons such that the number of possible combinations is astronomical, and that this may be sufficient to explain the diversity of mental representations, images and concepts. However, this appears to ignore the "curse of dimensionality" (Belman 1957). *If there is such a large space to search, how can adaptive pre-representations be produced sufficiently rapidly?* Changeux addresses this by suggesting that heuristics act on the search through pre-representations, notably, he allows recombination between neuronal assemblies, writing "this recombining activity would represent a 'generator of hypotheses', a mechanism of diversification essential for the geneses of pre-representations and subsequent selection of new concepts". However, a mechanism of recombination of function is *not* presented. In short, Changeux's mechanism struggles in the absence of copying.

Several mathematical models of the above arguments serve to test whether they could work in principle. Dehaene, Changeux and Nadal model the process of active storage of temporal sequences by matching with pre-representations. Note that there is no unit of evolution, for there is no entity that multiplies, and thus no inheritance. This fundamentally important limitation is admitted by the authors of the above model who write "an organism cannot learn more than is initially present in its pre-representations".

Later models incorporate stabilization of the configurations in a global workspace by *internal* reward and attention signals (Dehaene et al. 1998). In a model of the Stroop task, a global workspace is envisaged as having a repertoire of discrete activation patterns, only one of which can be active at once, and which can persist independent of inputs with some stability. This is meant to model persistent activity neurons in prefrontal cortex. These patterns constitute the selected entity (pre-representation), which "if negatively evaluated, or if attention fails, may be spontaneously and randomly replaced". Reward allows restructuring of the weights in the workspace. The improvement in performance depends on the global workspace having sufficient variation in patterns at the onset of the effortful task, perhaps with additional random variability, e.g. Dehaene and Changeux (1997) write that "in the

absence of specific inputs, prefrontal clusters activate with a fringe of variability, implementing a 'generator of diversity'". The underlying search algorithm is nothing more sophisticated than a random walk through pre-representation space, biased by reward.

Edelman's theory of neuronal group selection is very similar (Edelman 1987). It proposes that a primary repertoire of neuronal groups within the brain compete with each other for stimulus and reward resources. This results in selection of a secondary repertoire of behaviourally proficient groups (Izhikevich et al. 2004). The most modern version of Edelman's neuronal group selection is presented by Izhikevich et al. (2004). In Izhikevich's model, neurons are arranged in a recurrent network with delays in the passage of spikes between neurons and weights being modified by spike-time-dependent plasticity (STDP). Polychronous groups (PCGs) with stereotypical temporal firing patterns self-organize when a particular firing set of neurons is activated in a spatiotemporal pattern resulting in the convergence of spikes in downstream neurons. STDP subsequently reinforces these causal relationships. Because the same neuron can occur in many groups, and because delays produce an effectively infinite sized dynamical system (subject to temporal resolution constraints), the number of PCGs far exceeds the number of neurons in the network, allowing a very high memory capacity for stored PCGs. In a group of 1,000 neurons, approx 5,000 PCGs are found, with a distribution of sizes. If inputs are random, the *PCGs are formed and lost transiently*, over the course of minutes and hours. Structured input can stabilize certain PCGs. Izhikevich attempts to understand the functioning of polychronous groups within Edelman's framework of neuronal group selection (Izhikevich 2006). He proposes *two levels of selection*; selection at the neuronal level where spike-time-dependent plasticity (STDP) selects pathways to form PCGs, and selection between PCGs by inputs. Proposing two levels is rather a strange move when it seems sufficient to say that for a given set of spatiotemporal input patterns, there will be a set of PCGs that are attractors of the network dynamics. Nevertheless, Izhikevich has demonstrated that the number of these attractors is large, and that they are transient (given random inputs). PCGs are not units of selection in Izhikevich's model since he gives them no mechanism for replication. The most important fact to note is that *no mechanism is described showing how a beneficial trait of one PCG could be transmitted to another PCG.* The replication of connectivity described in Sect. 2.2 may be extended to the copying of polychronous groups, if it is possible to also copy delays. In conclusion, neither Edelman's nor Changeux's models include multiplication; therefore they do not describe a process of natural selection. The problem of transmission of a favourable trait of one group to non-group material is a ubiquitous feature of all versions of the theory of neuronal group selection.

The algorithms appear to reduce to variants of a reward-biased random walk. Stochastic hill-climbers at the neuronal level have been described by other authors, for example in the "Darwinian synapse" of Seung (2003). Michod summarizes the fact that in neuronal group selection, synaptic change rules replace replication as a mechanism of variability of the "unit of selection": there is correlation between the parental and offspring states of the same neuronal group even without multiplication

(Michod 1988). However, natural selection is a superior algorithm to biased stochastic search for at least the following reasons:

1. Replication allows a novel solution to be tested on a copy and not on itself, so that a harmful variant does not destroy the original solution.
2. Due to differential fitness and multiplication, search resources can be preferentially channelled to solutions that are the best so far.
3. Variability can be structured by the "copy + mutate + recombine" operators.

Whereas stochastic hill-climbing can explain performance in many simple tasks that can be solved by either random search or exhaustive search, e.g. the Stroop task (Dehaene et al. 1987, 1998), it is insufficient for solving tasks that require structured search in rugged and high-dimensional adaptive landscapes.

3.2 Reinforcement Learning Algorithms

There is evidence that the basal ganglia implement temporal difference (TD) learning using an actor-critic architecture (Houk et al. 2007). However, the differences between reinforcement learning and natural selection are subtle. TD methods are described as "using immediate rewards received by an agent to incrementally improve its estimated value function and thereby its policy" (Whiteson et al. 2007). That is, different state–action pairs exist at the outset of the task, and are assigned value by a value function as the task proceeds in an on-line manner. An action selection algorithm determines when to execute these state–action pairs as a function of their instantaneous value.

Let us attempt a re-description of this process from a natural selection perspective. Each state–action pair can be considered to be a replicator. The fitness of each replicator depends on the value that is assigned to it by the value function. The probability of application of a state–action pair during the task depends on the frequencies of the vector of actions available in that state. The differences between TD-learning and natural selection are made clear when one re-describes TD-learning in this way. The value function $Q\ (s,a)$ determines the fitness of each state–action pair, but unlike fitness functions defined by practitioners of genetic algorithms, the fitness function in reinforcement learning is designed to provide estimates of long-term value of taking action a in state s, at *each* time point in the task, not just at the end of the task. Sarsa and Q-learning are examples of value functions that update the value of either the action that was actually used at time t, or of the optimal action that could have been used at time t. Thus, RL methods differ from NS methods in implementing NS at a temporally finer-grained level by the use of a fitness function that can provide feedback at all time-steps. Furthermore, there is evidence that more complex value functions than mere reward prediction error are signalled by dopamine (Pennartz 1995; Redgrave et al. 1999). The second crucial difference is that RL methods lose the benefits of variability in state–action pairs that can arise from the imperfect replication operation. This is critical when the space of

state–action pairs is so large that it would be impossible to maintain all possible state–action pairs at the same time. The third difference is that on-line RL methods further constrain which state–action pairs are applied so as to balance exploration and exploitation. Such methods e.g. e-greedy selection, softmax selection and interval estimation can also be applied in evolutionary settings (Whiteson et al. 2007).

Whiteson et al. have experimented with using evolutionary approaches to evolving representations for value functions themselves and found that this improves the function of RL algorithms (Whiteson et al. 2007). Thus, NS can be used to evolve state–action pairs, *and* to evolve value functions. RL provides no explanation of how these representations originate.

3.3 Other Action Selection Approaches

The economist Brian Arthur writes that learning "can be viewed as dynamic competition among different hypotheses or beliefs or actions, with some reinforced and others weakened as fresh evidence and data are obtained" p. 133 (Arthur 1994). This is formalized in Holland-type classifier systems that consist of condition-action pairs that compete for application if their conditions are fulfilled. Issues with such systems are whether "the automaton does sufficient exploration to achieve asymptotic optimality" (p. 141) (Arthur 1994). The parameter space of annealing rates and rates of cooling can be fitted to human performance in two-choice bandit experiments.

Cisek proposes that neural populations in the fronto-parietal system represent entire distributions of potential movement parameters (Cisek 2006), i.e. cells encoding different motor parameters are activated differently, so that the population represents a probability density function over output space, in a variety of possible coordinate systems. Selection is by biasing signals from many other brain regions e.g. PFC and reward systems, and by intra-layer quenching competition. However, Cisek does not propose how such representations arise and are modified to suit the demands of novel tasks, a problem particularly relevant for cognitive rather than behavioral tasks.

The general problem with these models is that the space of actions is fixed and of low dimensionality. The system does not accumulate novel representational ability. All the variation exists at the outset, and the system simply restructures the probabilities with which particular actions are executed. The dimensionality problem also exists for Bayesian models where "the problem of efficiently representing and computing with probability density functions in high-dimensional spaces has been a barrier to developing efficient Bayesian computer vision algorithms" (Knill and Pouget 2004).

Natural selection allows a much larger space of solutions to be searched because there can be structured variability in condition-action space.

4 How is an Initial Population of Neuronal Representations Chosen?

How does a population of dynamical neuronal replicators required for a given search problem get initialized? This is a version of the frame problem (Pylyshn 1987), and remains a great mystery in neuroscience. According to Cisek, the prefrontal cortex (PFC) plays an important role in biasing action selection in posterior parietal cortex (Cisek 2006). Prefrontal cortex contains bistable neurons and recurrent circuits capable of persistent activity in the absence of sensory inputs (Miller and Cohen 2001; Rougier et al. 2005), and thus is suitable for working memory and dynamical neuronal replicators. Neuronal activity in PFC during a delayed response task is orthogonal between task phases, but correlated within task phases (Sigala et al. 2008). The neuronal replicator hypothesis predicts that structured variability consistent with structured search is to be found in a neuronal activity vector in the absence of stimulus presentation or behaviour, e.g. during sleep, and that this variability becomes more structured during development. Studies of spontaneous intrinsic dynamics support this view. Fox et al. write that "this alternative perspective suggests that the brain is active even in the absence of task, primarily driven by internal dynamics, with external events modulating rather than determining the dynamics of the system" (Fox et al. 2006). In the absence of task, stimuli or explicit attentional demands, they found spontaneous activity by fMRI in PFC and other regions. Llinás proposed that the "autorhythmic electrical properties of central neurons and their connectivity form the basis for an intrinsic functional coordinate system that provides internal context to sensory input" (Llinas 1988), and Fiser et al. report that "The correspondence between evoked neural activity and the structure of the input signal was weak in young animals, but systematically improved with age. This improvement was linked to a shift in the dynamics of spontaneous activity" (Fiser et al. 2004). A deeper understanding of how stimuli restructure the evolution of spontaneous activity is required to answer the question posed by this section.

5 On What Basis Are Neuronal Replicators Selected?

The question is of two parts, how is fitness assigned and what algorithm determines how inferior variants are replaced by superior ones? Midbrain dopamine systems signal the error in predicted reward (Izhikevich 2007; Oudeyer et al. 2007); however, recent evidence suggests that dopamine may also signal more complex value functions such as prediction error (Horvitz 2000). Friston and Stephen argue that the genetically specified goal of the brain is to reduce free energy which, given some simplifying assumptions, is equivalent to minimizing prediction error (Friston and Stephan 2007). However, it seems that such a function cannot explain why we seek novelty. In contrast to the claim of Ross Ashby, the role of the nervous system

is not merely to maintain essential variables within their viable range. If it were, we could not explain the tendency for exploration and curiosity. Extending the work of Friston and Stephen, Oudeyer et al. (2007) have proposed that the brain attempts to maximize learning progress, i.e. to produce continual increases in predictivity. Various other intrinsic value mechanisms have been proposed that involve structuring of input information (Lungarella and Sporns 2005), maximization of information flow in the sensori-motor loop (Klyubin et al. 2007), maximization of mutual information between the future and the past (Bialek et al. 2001), and these have been implemented to control robots (Der et al. 2008). From a different point of view, some co-evolutionary methods attempt to select for teachers on the basis of how well they discriminate between learners, suggesting that separate populations of neuronal replicators in the brain may have different functions, some units operating as fitness functions for other units that act as solutions or as predictors or perceptual agents (De Jong and Pollack 2003). Perhaps no other animal has the machinery to construct the rich internal value systems that humans possess, and which may be necessary for solving the credit assignment problem when explicit reward is delayed in time and when environments are non-Markovian (Hertz et al. 1991, p. 189).

Regarding the selection algorithm, neuronal evolutionary dynamics could turn out to be the best field of application of evolutionary graph theory (Lieberman et al. 2005). It has been shown that some topologies speed up, whereas others retard adaptive evolution. Figure 8 shows an example of favourable topologies (selection amplifiers). The brain could well influence the replacement topologies by gating, thereby realizing the most rewarding topologies.

Alternatively, Aunger proposes that a selection algorithm on neuromemes is implemented by i) neural synchrony (Singer 1999) and may correspond to the phenomena of selective attention (Steinmetz et al. 2000). There is some evidence that the decay of synchrony is an indicator of a change in the "train of thought" (Rodriguez et al. 1999).

6 How Can Variability Be Structured?

An unstructured heuristic search is highly inefficient in complex problem domains (Richards 1977). A random search in any optimization problem with multiple interdependencies (Watson 2006), for example in Ross Ashby's homeostat becomes extremely slow to reach a stable state given a large and highly connected network (Ashby 1960). Replicators that structure their exploration distributions are ubiquitous in all systems capable of open-ended evolution (Bedau 1998). The evolution of evolvability (Pigliucci 2008) can occur by hitchhiking of neutral variability-structuring mutations that are selected because they shape the exploration distribution of variants (Kirchner and Gerhart 1998; Toussaint 2003). If there is variation in variability then selection can act on that variation (Wagner and Altenberg 1996). Structured search is such a deep feature of conscious thought that we take it for granted.

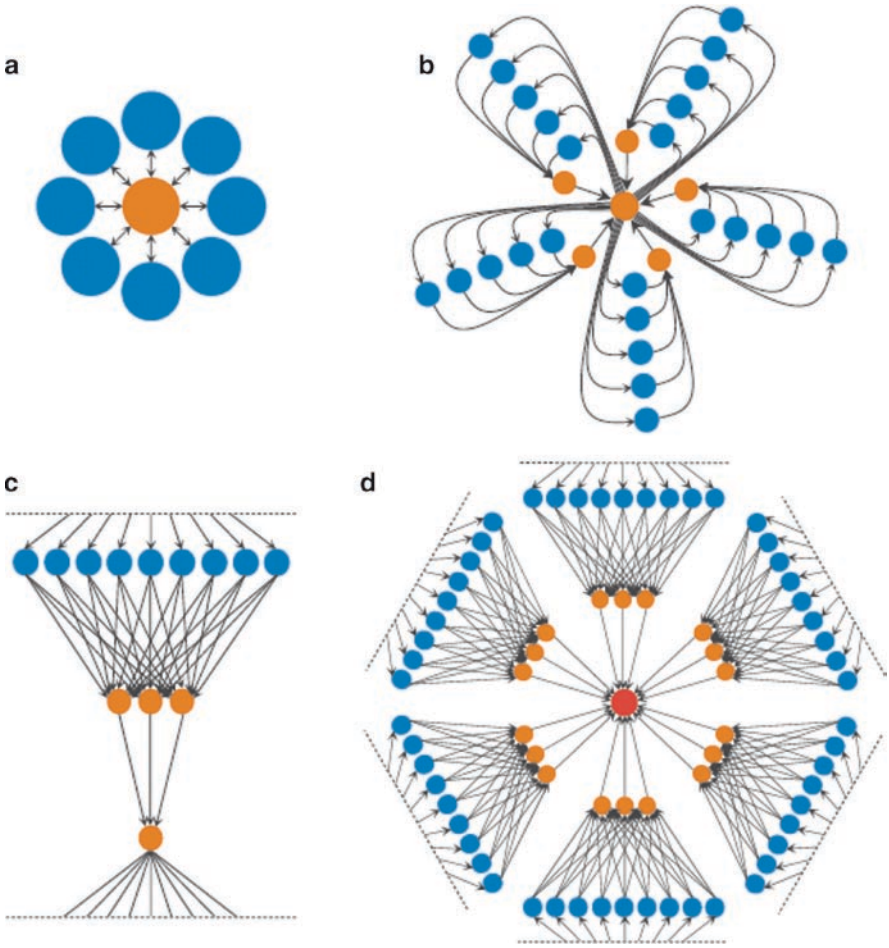

Fig. 8 Selection amplifiers from Lieberman et al. (2005). They may be implemented in nerve tissue. Vertices that change often, due to replacement from the neighbours, are coloured in orange. In the present context each vertex can be a neuron or neuronal group that can inherit its state from its upstream neighbours and pass on its state to the downstream neighbours. Neuronal evolution would be evolution on graphs

Our thoughts are neither random nor exhaustive. The proposal here is that a unifying principle underlies structured thought and structured evolutionary search.

Although the reduction of mutation rate (increase of accuracy) is invariably good up to a certain level, evolution would come to a halt with perfect accuracy (which, by the way, cannot be achieved anyway). It is important to quote the results by Jones et al. (2007) who have demonstrated in a quantitative genetics model that the mutation matrix **M** can align itself with the adaptive landscape which favours

evolvability (the production of advantageous variation). The result holds for stabilizing selection but it is suspected that it generalizes to systems under directional selection. Obviously, in an evolutionary system which would be used for real-time problem-solving (as it may happen in the nervous system) the **M** matrix and its effect selection could play a crucial role also. In particular, a neat mechanism going against the strictly "blind variation" of Campbell (1974) can be envisaged, and it could be more important in the present context than for populations of organisms.

This mechanism involves the use of Hebbian learning to structure the copy operation. Imagine that, instead of limiting between-layer connections to a topographic map, one starts with a strong one-to-one topographic map and allows all-to-all Hebbian connections to develop once a local (or global) optimum has been reached. After the optimum activity vector has been obtained and is present in both parent and child layer, Hebbian learning is permitted between these two vectors (for all synapses except for the original one-to-one topographic connections). If the activity vectors are then reset and a new evolutionary run is started, then copying will be biased by the Hebbian learning that took place in previous evolutionary runs. An active neuron in the parental layer will not only tend to activate the corresponding one-to-one topographic neuron, but also other neurons in the offspring layer that were previously active when the optimal solution had been found. Oja's rule is used to control the Hebbian between-layer synapses. We have found that if Hebbian learning is permitted, then later evolutionary searches converge faster, because they learn from previous evolutionary searches. Richard Watson and colleagues in Southampton have described a set of search problems that are particularly well suited to neuronal copying biased by Hebbian learning (Watson et al. 1998; Watson 2006) and have proposed that "symbiotic evolution" can effectively solve these problems in a wide range of domains, one of which is neuronal replication (Watson et al. 2009).

7 Cognitive Roles for Neuronal Replicators

The copying mechanisms described depend on topographic maps to act as the neuronal equivalent of h-bonds. This adds to the possible role of topographic maps in cognition (Thivierge and Marcus 2007). We make the empirical prediction that our neuronal replicators should be found perpendicular to topographic maps outside sensory areas, for example perpendicular to CA1 hippocampo-entorhinal projections and nigrostriatal projections. Three applications in cognition are discussed.

7.1 Causal Inference by Topology Copying

The mechanism of neuronal copying is a neuronal implementation of causal inference (Glymour 2003). The capacity of STDP to capture temporal relations consistent with causality rather than just correlations has been described by several

authors (Abbott and Nelson 2000; Bi and Poo 2001; Florian 2007; Gerstner and Kistler 2002). However, to our knowledge, STDP has until now not been used in an algorithm to explicitly infer whole causal networks. Considerable attention has been paid recently to the capacity of animals such as New Caledonian crows (Weir et al. 2002), rats (Blaisdell 2006), non-human apes (Blaisdell 2006), children (Gopnik and Schulz 2004) and human adults (Gopnik and Schulz 2004) to undertake causal reasoning tasks, i.e. tasks in which good performance cannot be well explained by pair-wise associative learning alone. A Bayesian account can be given of performance in some tasks (Orbán et al. 2008). Another approach is to give a constraint-based reasoning account that involves graph operations on a Bayes Net and interventions to discover conditional independencies between nodes (Pearl 2000). Recent work reveals that humans use temporal order, intervention, and covariation to produce generative models of external events (Lagnado et al. 2007). The STDP-based copying algorithm we describe does the same kind of thing; it infers a dynamical causal model from a set of spike trains that arise from an underlying and invisible causal graph (another neuronal network). If instead this set of spike trains arises from a sensory system, in which the underlying causal graph exists in the outside environment, then the same inference mechanism can be used to produce a neuronal generative model (Dayan et al. 1995) of these external stimuli. Such forward models feature in influential theories of cognition (Churchland 2002) in the form of emulators (Craik 1943; Grush 2004).

7.2 Heuristic Search in Insight Problems and Working Memory Tasks

Selective attention in complex tasks (Desimone and Duncan 1995) and solving insight problems (Chronicle et al. 2004) may require a mechanism for structuring search. Exhaustive or random search may be too slow, and hill-climbing may be inefficient due to local-optima. “An important feature of future research will be to identify neural mechanisms that implement more sophisticated forms of search” (Rougier et al. 2005). Indeed we do not yet know what neural mechanisms underlie human creativity in problems such as the nine-dot problem: “Draw four continuous straight lines, connecting all the dots without lifting your pencil from the paper” (MacGregor et al. 2001).

Solutions to insight problems may exist on rugged fitness landscapes (Perkins 1995), i.e. there may be problem interdependencies (Watson et al. 1998). Intrinsic value functions must be created and used to guide search (MacGregor et al. 2001) (Chronicle et al. 2004; Simon and Reed 1976). Subconscious processes are also clearly involved as evidenced by the fact that sleep doubles the rate at which explicit knowledge is gained about a hidden correlation in a stimulus-response task (Wagner et al. 2004), and that one is unaware of why a particular solution came to mind (Sternberg and Davidson 1995).

The heuristic search hypothesis of (Newell and Simon 1976) assumes that "solutions to problems are represented as symbol structures. A physical symbol system exercises its intelligence in problem solving by search – that is, by generating and progressively modifying symbol structures until it produces a solution structure". The physical symbol system hypothesis has been heavily criticized (Brooks 1990). But we note there appears to be a parallel in the requirement for "symbols" both in the brain and in organismal natural selection as became evident in the problem of how to maintain information and diversity by blending inheritance, (Gould 2002, p. 622). The only solution was to allow symbolic, i.e. particulate Mendelian inheritance (Fisher 1930). Symbols seem to be a crucial requirement for natural selection with unlimited heredity, irrespective of its implementation. This indirectly confirms the physical symbol system hypothesis in the brain.

7.3 Memory Consolidation

To store something in long-term memory is to stop its modification, i.e. to remove it from working memory, a process of rapid activity-dependent search, and to embed it in patterns of (protein-dependent) neuronal connectivity where it can be utilized (retrieved) either consciously as in an episodic memory, used to structure thought as in semantic memory, or unconsciously to structure behavior as in procedural memory. The loops between the medial temporal cortex (containing the hippocampus) and the neocortex have been implicated in memory consolidation and reconsolidation. The "integrative function [of the hippocampus] is taken over by the medial prefrontal cortex" at least for semantic memories (Frankland and Bontempi 2006). One possibility is that topology copying is involved in this transfer of function from hippocampus to neocortex. Secondly, a process of neuronal topology evolution may play a role in the multiple trace theory of consolidation and reconsolidation. Multiple trace theory (MTT) suggests that complex interactions between the hippocampus and the neocortex including the prefrontal cortex are involved in consolidation and recall of episodic memories (Nadel and Moscovitch 1997). The MTT proposes that a new hippocampus-dependent memory trace is created whenever an episode is retrieved. Memory traces "decay (i.e. disappear) and can replicate", in a model that explains some properties of the loss of memory as a function of lesion size (Nadel et al. 2000). However, they provide no description of the internal structure or dynamics of a memory trace that allows it to replicate.

Some of the features of the hippocampus that may allow memory trace formation include synaptic plasticity, formation of new synapses on dendrites (via stabilization of filopodia in a calcium-dependent manner (Lohmann and Bonhoeffer 2008)) and unsilencing of silent (AMPA) synapses (Xiao et al. 2004) particularly in conjunction with adult neurogenesis, which is the formation of new (immature) neurons in the dentate gyrus region of the hippocampus (Cameron and McKay 2001). The unsilencing of silent synapses, would likewise enable new or modified traces to be formed without interfering with the existing traces. The formation of

new synapses in a manner based on calcium-dependent signalling selecting the survival of a potential synaptic partner formed by a filopodia (Lohmann and Bonhoeffer 2008) could be envisaged to allow new contacts to be made between neurons in the same layer.

8 Conclusions

The above models and considerations are very preliminary; however, we hope they will inspire further experimental work in an attempt to detect neuronal replicators. We are unlikely to understand the human brain until we have produced machines that exhibit comparable levels of plastic adaptation, at least to an autistic level. The neuronal replicator hypothesis should be tested in both the experimental and engineering domains. The crux of our belief in the hypothesis lies in the fact that only natural selection is known to account for cumulative open-ended adaptation of the type observed in cognition. Sufficient reason to examine the neuronal replicator hypothesis further comes from the biological plausibility of neuronal replication mechanisms in conjunction with the need to explain structured search in cognition.

Acknowledgement Thanks to a Career Development Fellowship at the MRC National Institute for Medical Research, Mill Hill, London, and to a Marie Curie Inter-European Grant to work at Collegium Budapest, Hungary. Partial support of this work has generously been provided by the Hungarian National Office for Research and Technology (NAP 2005/KCKHA005), by the Hungarian Scientific Research Fund (OTKA, NK73047) and the eFlux FET-OPEN project (225167). The work has also been supported by the COST Action CM0703 on Systems chemistry. We thank Vinod Goel, Mauro Santos, Richard Goldstein and Michael Öllinger for their helpful suggestions.

References

Abbott L, Nelson SB (2000) Synaptic plasticity: taming the beast. Nat Neurosci Suppl 3:1178–1183

Abeles M (1982) Local cortical circuits: an electrophysiological study. Springer, Berlin

Abeles M (1991) Corticonics: neural circuits of the cerebral cortex. Cambridge University Press, New York

Abeles M, Gat I (2001) Detecting precise firing sequences in experimental data. J Neurosci Methods 107:141–154

Adams P (1998) Hebb and Darwin. J Theor Biol 195(4):419–438

Arthur WB (1994) Increasing returns and path dependence in the economy. University of Michigan Press, Michigan

Ashby R (1960) Design for a brain. Wiley, New York

Aunger R (2002) The electric meme: a new theory of how we think. Free, New York

Baldwin MJ (1898) On selective thinking. Psychol Rev 5(1):4

Baldwin MJ (1909) The influence of Darwin on theory of knowledge and philosophy. Psychol Rev 16:207–218

Barbrook AC, Howe CJ et al (1998) The phylogency of the Canterbury tales. Nature 394:839
Bedau MA (1998) For puzzles about life. Artif Life 4(2):125–140
Belman RE (1957) Dynamic programming. Princeton University Press, Princeton, NJ
Beyer H-G (2001) The theory of evolution strategies. Springer, Berlin
Bi G-q, Poo M-m (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 24:139–166
Bialek W, Nemenman I, et al (2001) Predictability, complexity and learning. Neural Comput 13:2409
Blaisdell A (2006) Causal reasoning in rats. Science 311(5763):1020–1022
Boden M (2006) Mind as machine: a history of cognitive science. Oxford University Press, Oxford
Boyd R, Richerson PJ (2005) The origin and evolution of cultures. Oxford University Press, Oxford
Brooks R (1990) Elephants don't play chess. Robot Auton Syst 6:3–15
Calvin WH (1987) The brain as a Darwin Machine. Nature 330:33–34
Calvin WH (1996) The cerebral code. MIT, Cambridge, MA
Cameron HA, McKay RD (2001) Adult neurogenesis produces a large pool of new granule cells in the dentate gyrus. J Comput Neurol 435:406–417
Campbell DT (1974) The philosophy of Karl. R. Popper. In Schillpp PA (ed) Evolutionary epistemology. University of Chicago Press, Chicago, pp 412–463
Cateau H, Fukai T (2001) Fokker-Planck approach to the pulse packet propagation in synfire chain, Neural Networks 14:675–685
Changeux JP, Courrege P et al (1973) A theory of the epigenesis of neuronal networks by selective stabilization of synapses. Proc Natl Acad Sci U S A 70:2974–2978
Changeux JP (1985) Neuronal man: the biology of mind. Princeton University Press, Princeton
Chronicle EP, MacGregor JM et al (2004) What makes an insight problem? The roles of heuristics, goal conception, and solution recording in knowledge-lean problems. J Exp Psychol Learn Mem Cogn 30(1):14–27
Churchland P (2002) Brain-wise: studies in neurophilosophy. Bradford Book, Cambridge, MA
Cisek P (2006) Integrated neural processes for defining potential actions and deciding between them: a computational model. J Neurosci 26(38):9761–9770
Cooper W (2001) The evolution of reason: logic as a branch of biology. Cambridge University Press, Cambridge
Craik K (1943) The nature of explanation. Cambridge University Press, Cambridge, UK
Crick F, Mitchison G (1995) REM sleep and neural nets. Behav Brain Res 69:147–155
Crick FHC (1989) Neuronal Edelmanism. Trends Neurosci 12:240–248
Crick FHC (1990) Reply. Trends Neurosci 13:13–14
Dawkins R (1982) The extended phenotype: the gene as the unit of selection. Freeman, Oxford
Dayan P, Abbott L (2001) Theoretical neuroscience: computational and mathematical modeling of neural systems. MIT, Cambridge, MA
Dayan P, Hinton GE et al (1995) The Helmholtz machine. Neural Comput 7:1022–1037
De Jong E, Pollack JB (2003) Learning the ideal evaluation function. LNCS GECCO 2723:203
Dehaene S, Changeux JP (1997) A hierarchical neuronal network for planning behavior. Proc Natl Acad Sci U S A 94(24):13293–13298
Dehaene S, Changeux JP et al (1987) Neural networks that learn temporal sequences by selection. Proc Natl Acad Sci U S A 84(9):2727–2731
Dehaene S, Kerszberg M et al (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci U S A 95(24):14529–14534
Dennett DC (1981) Brainstorms. MIT, Cambridge, MA
Dennett DC (1995) Darwin's dangerous idea. Simon & Schuster, New York
Der R, Guttler F, et al (2008) Predictive information and emergent cooperativity in a chain of mobile robots. Artificial Life, Southampton, UK, pp 166–172
Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193–222

Edelman GM (1987) Neural Darwinism. The theory of neuronal group selection. Basic Books, New York
Edelman GM (1994) The evolution of somatic selection: the antibody tale. Genetics 138:975–981
Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58(10):465–523
Fernando C, Karishma KK, Szathmáry E (2008) Copying and Evolution of Neuronal Topology. PLoS ONE 3(11): e3775. doi:10.1371/journal.pone.0003775.
Fiser J, Chiu C et al (2004) Small modulation of ongoing cortical dynamics by sensory input during natural vision. Nature 431:573–578
Fisher RA (1930) The genetical theory of natural selection. Clarendon, London
Florian RV (2007) Reinforcement learning through modulation of spike-time-dependent synaptic plasticity. Neural Comput 19:1468–1502
Fodor JA, Pylyshyn ZW (1988) Connectionism and cognitive architecture: a critical analysis. Cognition 28:3–71
Fox MD, Corbetta M et al (2006) Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. Proc Natl Acad Sci U S A 103:10046–10051
Frankland PW, Bontempi B (2006) Fast track to the medial prefrontal cortex. Proc Natl Acad Sci 103(3):509–510
Friston KJ, Stephan KE (2007) Free-energy and the brain. Synthese 159:417–458
Gánti T (2003) The principles of life. Oxford University Press, Oxford, UK
Gerstner W, Kistler WM (2002) Mathematical formulations of Hebbian learning. Biol Cybern 87:404–415
Glymour C (2003) Learning, prediction and causal Bayes nets. Trends Cogn Sci 7(1):43–48
Gopnik A, Schulz L (2004) Mechanisms of theory formation in young children. Trends Cogn Sci 8(8):371–377
Gould SJ (2002) The structure of evolutionary theory. The Belknap Press of Hardvard University Press, Cambridge, MA
Greisemer JR (2000) Development, culture, and the units of inheritance. Philos Sci 67:348–368
Grush R (2004) The emulation theory of representation: motor control, imagery, and perception. Behav Brain Sci 27:377–442
Hadamard J (1945) The psychology of invention in the mathematical field. Dover, New York
Harvey I (2008) Misrepresentations. In Bullock JNS, Watson RA, Bedau MA (eds) Proceedings of the Eleventh International Conference on Artificial Life. Winchester, UK. MIT, Cambridge, MA, pp 227–233
Hasselmo ME (2006) The role of acetylcholine in learning and memory. Curr Opin Neurobiol 16:710–715
Hebb DO (1949) The organization of behaviour. Wiley, New York
Hertz J, Krogh A, et al (1991) Introduction to the theory of neural computation. Westview, Tennessee
Horvitz J-C (2000) Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience 96(4):651–656
Houk JC, Bastianen C et al (2007) Action selection and refinement in subcortical loops through basal ganglia and cerebellum. Philos Trans R Soc B 29:1573–1583
Izhikevich EM (2006) Polychronization: computation with spikes. Neural Comput 18(2):245–282
Izhikevich EM (2007) Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb Cortex 17:2443–2452
Izhikevich EM, Gally JA et al (2004) Spike-timing dynamics of neuronal groups. Cereb Cortex 14(8):933–944
Jablonka E, Lamb MJ (2005) Evolution in four dimensions: genetic, epigenetic, behavioral, and symbolic variation in the history of life. Bradford Books, Bradford, UK
James W (1890) The principles of psychology. Dover, New York
Jones AG, Arnold SJ et al (2007) The mutation matrix and the evolution of evolvability. Evolution 61:727–745

Jones BL, Enns RH et al (1976) On the theory of selection of coupled macromolecular systems. Bull Math Biol 38:15–28
Kirchner M, Gerhart J (1998) Evolvability. Proc Natl Acad Sci U S A 95:8420–8427
Klyubin AS, Polani D et al (2007) Representations of space and time in the maximization of information flow in the perception-action loop. Neural Comput 19:2387–2432
Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27(12):712–719
Lagnado DA, Waldmann MR, et al (2007) Beyond covariation: cues to causal structure. In Gopnik A, Schulz L (eds) Causal learning: psychology, philosophy, and computation. Oxford University Press, Oxford, UK, pp 154–172
Lenski RE, Ofria C et al (1999) Genomic complexity, robustness, and genetic interactions in digital organisms. Nature 400:661–664
Lieberman E, Hauert C et al (2005) Evolutionary dynamics on graphs. Nature 433:312–316
Llinas RR (1988) The intrinsic electrophysiological properties of mammalian neurons: insights into central nervous system function. Science 242:1654–1664
Lohmann C, Bonhoeffer T (2008) A role for local calcium signaling in rapid synaptic partner selection by dendritic filopodia. Neuron 59:253–260
Lungarella M, Sporns O (2005) Information self-structuring: key principle for learning and development. IEEE Int Conf Dev Learn 2005:25–30
MacGregor JM, Ormerod TC et al (2001) Information processing and insight: a process model of performance on the nine-dot and related problems. J Exp Psychol Learn Mem Cogn 27(1):176–201
Markram H, Lubke J et al (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275:213–215
Marr D (1983) Vision: a computational investigation into the human representation and processing of visual information. Freeman, Oxford
Maynard Smith J (1986) The problems of biology. Oxford University Press, Oxford
Maynard Smith J, Szathmáry E (1998) The major transitions in evolution. Oxford University Press, Oxford
Michod RE (1988) Darwinian selection in the brain. Evolution 43(3):694–696
Miller EK, Cohen DJ (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 2001(24):167–202
Mitchell M (1996) An introduction to genetic algorithms. MIT, Cambridge, MA
Monod, J (1971) Chance and necessity: an essay on the natural philosophy of modern biology. Knopf, New York
Muller HJ (1966) The gene material as the initiator and organizing basis of life. Am Nat 100:493–517
Nadel L, Moscovitch M (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. Curr Opin Neurobiol 7:217–227
Nadel L, Samsonovich A et al (2000) Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. Hippocampus 10:352–368
Newell A, Simon HA (1972) Human problem solving. Prentice-Hall, Englewood Cliffs, NJ
Newell A, Simon HA (1976) Computer science as empirical inquiry: symbols and search. Commun Assoc Comput Mach 19(3):113–126
Oja E (1982) Simplified neuron model as a principal component analyzer. J Math Biol 15(3):267–273
Okasha S (2006) Evolution and the levels of selection. Oxford University Press, Oxford
Orbán G, Fiser J et al (2008) Bayesian learning of visual chunks by human observers. Proc Natl Acad Sci U S A 105(7):2745–2750
Oudeyer P-Y, Kaplan F et al (2007) Intrinsic motivation systems for autonomoys mental development. IEEE Trans Evol Comput 11(2):265–286
Pearl J (2000) Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, UK

Pennartz CM (1995) The ascending neuromodulatory systems in learning by reinforcement: comparing computational conjectures with experimental findings. Brain Res Rev 21:219–245
Perkins DN (1995) Insight in minds and genes. In: Sternberg RJ, Davidson JE (eds) The nature of insight. MIT, Cambridge, MA
Pigliucci M (2008) Is evolvability evolvable? Nat Rev Genet 9:75–82
Price GR (1970) Selection and covariance. Nature 227:520–521
Pylyshn ZW (ed) (1987) The robot's dilemma: the frame problem in artificial intelligence. Norwood, NJ, Ablex
Redgrave P, Prescott TJ et al (1999) Is the short-latency dopamine response too short to signal reward error? Trends Neurosci 22:146–151
Richards RJ (1977) The natural selection model of conceptual evolution. Philos Sci 44(3):494–501
Rodriguez E, George N et al (1999) Perception's shadow: long distance synchronization of human brain activity. Nature 397:157–161
Rougier NP, Noelle DC et al (2005) Prefrontal cortex and flexible cognitive control: rules without symbols. Proc Natl Acad Sci U S A 102:7338–7343
Seung SH (2003) Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron 40:1063–1973
Sigala N, Kusunoki M et al (2008) Hierarchical coding for sequential task events in the monkey prefrontal cortex. Proc Natl Acad Sci U S A 105:11969–11974
Simon HA (1996) The sciences of the artificial. MIT, Cambridge, MA
Simon HA, Reed SK (1976) Modeling strategy shifts in a problem-solving task. Cogn Psychol 8:86–97
Singer W (1999) Neuronal synhrony: a versatile code for the definition of relations? Neuron 24(1):49–65
Song S, Abbott L (2001) Cortical development and remapping through spike timing-dependent plasticity. Neuron 32:339–350
Steinmetz PN, Roy A et al (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature 404:187–189
Sternberg RJ, Davidson JE (eds) (1995) The nature of insight. MIT, Cambridge, MA
Thivierge J-P, Marcus GF (2007) The topographic brain: from neural connectivity to cognition. Trends Neurosci 30(6):251–259
Thompson CJ, McBride JL (1974) On eigen's theory of the self-organization of matter and the evolution of biological macromolecules. Math Biosci 21:127–142
Toussaint M (2003) The evolution of genetic representations and modular adaptation. ND 04, Bochum, Germany
Van Rossum MC, Bi G et al (2000) Stable hebbian learning from spike-timing dependent plasticity. J Neurosci 20:8812–8821
Wagner GP, Altenberg L (1996) Complex adaptations and evolution of evolvability. Evolution 50:329–347
Wagner U, Gais S et al (2004) Sleep inspires insight. Nature 427:352–355
Watson RA, Buckley CL, Mills R (2009) The Effect of Hebbian Learning on Optimisation in Hopfield Networks. Technical Report, ECS, University of Southampton
Watson RA (2006) Compositional evolution: the impact of sex, symbiosis, and modularity on the gradualist framework of evolution. MIT, NA
Watson RA, Hornby GS, et al (1998) Modelling building-block interdependency. Proceedings of Fifth International Conference/PPSN V, Springer, Berlin
Weir AAS, Chappell J et al (2002) Shaping of hooks in New Caledonian crows. Science 297:981
Whiteson S, Taylor ME et al (2007) Empirical studies in action selection with reinforcement learning. Adapt Behav 15:33–50
Willshaw D, von der Malsburg C (1976) How patterned neural connections can be set up by self-organisation. Proc R Soc Lond B 194:431–445

Xiao M-Y, Wasling P et al (2004) Creaction of AMPA-silent synapses in the neonatal hippocampus. Nat Neurosci 7:236–243
Young JM, Waleszczyk WJ et al (2007) Cortical reorganization consistent with spike timeing but not correlation-dependent plasticity. Nat Neurosci 10(7):887–895

Value and Self-Referential Control: Necessary Ingredients for the Autonomous Development of Flexible Intelligence

Olaf Sporns and Edgar Körner

Abstract Internal evaluations, motivations, contexts, goals and plans play a crucial role in the control of behavior and cognition. In this review, we argue that capable neurorobotic systems need to incorporate a flexible and dynamic architecture that supports the self-organization of value and knowledge representation by means of self-referential control. Such architecture forms an indispensable basis for the autonomous development of both simple adaptive behavior and higher cognition. We provide a brief review of empirical and theoretical work addressed in this area, outline a set of design principles for a self-organizing and open-ended knowledge architecture, and provide a strategy for its implementation in intelligent systems.

1 Introduction

It has been stated by many investigators in neuroscience and cognitive science that many aspects of human cognition, including high-level cognitive functions involving conscious thought processes, continuously make reference to the internal state of the body, and thus to the fundamental and evolutionarily ancient value systems of the organism. As thought emerges from the autonomous development of knowledge representations, it remains bound to the brain's value systems and subject to self-referential control, and it is this link with the body that contributes to the definition of the "self" as a distinct cognitive and psychological entity. In their simplest forms, value systems communicate the saliency of sensory events in the environment and they thus play important roles in organizing simple behaviors into adaptive actions.

O. Sporns (✉)
Department of Psychological and Brain Sciences, Indiana University,
Bloomington, Indiana, 47405, USA
e-mail: osporns@indiana.edu

E. Körner
Honda Research Institute Europe, Carl-Legien Strasse 30, Offenbach/Main, 63073, Germany
e-mail: edgar.koerner@honda-ri.de

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_21, © Springer-Verlag Berlin Heidelberg 2010

We argue that value systems and the mechanisms of self-referential control with which they are intimately connected are essential ingredients for shaping and organizing the complex activities of the brain into meaningful thoughts and actions.

The architecture of biological nervous systems is extraordinarily complex, with distributed networks spanning multiple levels of organization that interact, develop over time and generate a continuous sequence of perceptual and cognitive states underlying behavior. Behavior itself is determined by the dynamic interplay of external and internal causes. External causes include sensory stimuli that originate from objects and events in the environment and are relayed to the nervous system through sense organs. Such external causes are complemented by internal causes that include the motivational state of the organism, prior knowledge about a stimulus or task domain, recalled memories providing context for current action, expertise, attention, as well as behavioral goals and plans. The importance of value (or valuation, or evaluation), motivation and emotion in shaping behavioral goals, decision making and learning in humans has been underscored by numerous behavioral and cognitive neuroscience studies (Rolls 1999; LeDoux 1996; Toates 1986; Dolan 2002; Berridge 2004; Sugrue et al. 2005). The central role of an organism's internal state for its neural and cognitive development, for its ability to engage in organized action and for its capacity to shape its own development in the course of open-ended exploration is now widely recognized (Dalgleish 2004; Arbib and Fellous 2004).

The emerging field of neurorobotics (or embodied artificial intelligence) aims at combining mechanisms and concepts from neuroscience with the design of robotic platforms (Chiel and Beer 1997; Lungarella et al. 2003; Sporns 2005; Reeke et al. 2005; Seth et al. 2005). A major research goal is for neurorobotic systems to become truly autonomous and adaptive, essentially at a level equal to that of biological organisms, specifically humans (Asada et al. 2001). We suggest that this research goal can be furthered by identifying basic principles of autonomous development in organisms (including humans) and by implementing these principles in robots (Asada et al. 2001; Weng et al. 2001; Weng and Zeng 2005).

In this chapter, we identify some of the structural, architectural and dynamic principles that may support the development of a flexible intelligent system. What are the minimal components that are needed for a self-organizing and self-referential model of value and knowledge representation? What are the key neural structures (architecture), sensory capacities (modalities, resolution), dynamic capabilities (learning, plasticity), and environmental factors (richness, variability) that are required? Why does such a model have to be self-organizing and self-referential?

The paper is divided into four sections that discuss two candidate principles for autonomous development: value and self-referential control. First, we define and argue for the importance of prior (innate) knowledge to "jump-start" autonomous development. Then, we discuss the basic need for progressive modification of innate knowledge to allow for the incorporation of environmental features that are individual and historic. We then highlight the crucial role of exploratory behavior

in generating variability as well as novelty and surprise. Finally, we provide a basis for the development of self-referential control through the subjective linkage of knowledge representations to substrates of valuation and internal state.

2 Value and Self-Referential Control

What are the key principles of a control architecture that must be implemented in order to make a robotic system capable of creating and sustaining self-organized knowledge representations? Our answer to this question is built around two main principles: 'value' (Reeke et al. 1990; Friston et al. 1994; Sporns et al. 2000; Huang and Weng 2002) and 'self-referential control' (Körner and Matsumoto 2002).

A central design principle responsible for linking information sampled from the environment (external inputs and stimuli) to the internal needs and goals of a behaving system is the principle of value. Value has been conceptualized as imposing important biases on the outcome of interactions with the environment and generating neural signals that are essential to reflect the global evaluation of recent behavior (Reeke et al. 1990). Value systems are an integral part of the neural control architecture corresponding to diffuse and neuromodulatory ascending systems of the brain. Value systems generate global signals that are broadcast to widespread areas of the brain and that are internally derived by the behaving system after actual behavior has occurred (Reeke et al. 1990; Friston et al. 1994; Sporns et al. 2000). Value systems are inherently multilevel and distributed, i.e., their effects are regionally differentiated and specific.

Self-referential control is essential for the emergence of autonomous intelligent systems (Körner and Matsumoto 2002). An essential point about self-referential architectures is that their structure reflects not only stored knowledge, memory items, or processed information, but the structure itself is instrumental in guiding the acquisition of future knowledge in a fully autonomous mode. This acquisition is carried out in the absence of specific external control signals that instruct the architecture and without an "end point" that corresponds to final convergence onto an optimal end state. The architecture itself embodies an algorithm for how to acquire knowledge and interpret it on the basis of its own internal state. Autonomous acquisition of new knowledge and integration of this knowledge into an existing relational structure ensures the continued growth of a consistent subjective (i.e., self-referential) knowledge representation.

Value and self-referential control are basic principles that are firmly grounded in neurobiology. Modulatory neurotransmitter systems, including dopamine, serotonin, and acetylcholine, possess many of the structural and functional properties of value systems including their capacity to acquire new response characteristics in the course of experience (see below). Throughout the nervous system, neuromodulators are involved in the regulation of neuronal excitability and plasticity to effects on gene expression and structural modifications in neural circuits. They also regulate brain circuits that control behaviors, including the processing of

rewarding and aversive stimuli, and modulation of cortical areas involved in working memory and executive control. Neuromodulatory neurons and transmitters are found in virtually all vertebrate (Hasselmo et al. 2002) and several invertebrate species (Hammer 1993, 1997; Birmingham and Tauck 2003). In computational models, neuromodulation has been implemented as effects on neuronal response functions, on learning rates or synaptic efficacies, or other model parameters (Servan-Schreiber et al. 1990; Fellous and Linster 1998; Doya 2000; Hasselmo 1995). Broadly, neuromodulatory effects can be categorized as resulting in changes of synaptic efficacies that may lead to persistent alterations of behavioral patterns and as changes of the response properties and local dynamics of neurons and neuronal circuits that may lead to alterations in information processing and cognitive function.

Value and self-referential control must promote autonomous development of the organism or robot, starting from simple sensory-motor mappings and progressing to increasingly refined evaluation strategies that depend on actual experience. How can we plot a strategy for the implementation of such autonomous value systems?

2.1 Elements of Innate Knowledge

What do we mean by "autonomous"? We suggest that autonomy always involves the capability, *of the system itself*, to select an appropriate behavioral reaction given a specific sensory situation. This selection involves attaching a behaviorally defined value to a specific configuration of sensory inputs. For example, simple organisms respond to trigger features (sign stimuli) in their sensory signal space much like reflex automata, i.e., by producing stereotypic behaviors, such as attack or escape, mating or grooming (Tinbergen 1951). The relation between such a stereotypic behavior or fixed action patterns and the respective trigger stimulus is encoded by genetic information which has been acquired in an evolutionary optimization process. Thus, the basic mapping of sensory inputs to specific classes of behavior constitutes innate information that encodes, from an evolutionary perspective, an optimal strategy for responding to a dynamic sensory environment while maximizing chances for survival.

From the point of view of the system's organization, the genetically determined mapping of trigger features to prototypic behaviors defines a coarse but fundamental metric of value assignment. It is important to note that in this context value is defined from a subjective point of view, with the survival of the subject being at the core of the value setting. In that respect, despite being imprinted, the innate setting of the fundamental value metric reflects a subjective evaluation of the sensory space, which is inherently consistent with and, hence, can be confirmed and increasingly amplified by the sensory experience of the subject. The metric itself is innate, but not defined from an outside point of view, not based on an evaluation which the system by itself cannot verify and extend based on its own sensory experience during interaction with the environment. Consistency of the innately

defined value metric with the (also innate) reinforcement learning system architecture provides the outline of a consistent "belief" structure, which enables the system both to behave properly from the beginning and to extend and diversify (as described in the next section) the innate value metric according to its experience in interacting with the environment.

No organism or robotic system can undergo an extensive process of self-organization (developmental or evolutionary) without building on innate or preexisting structures and processes that can serve as starting points and boundary constraints for all future developmental trajectories. There is overwhelming evidence that brains and bodies are highly structured and differentiated right from the very beginning of development (Edelman 1988). This structure reflects genetic and evolutionary history, which molds organisms according to very basic and essentially invariant physical and informational properties of their environments. In the case of robots, evolution is replaced by a set of design specifications.

We may distinguish three main dimensions of prior or innate knowledge: (1) Morphology. The structure of the physical body and of the neural architecture is crucial for enabling basic control of simple behaviors. Morphology comprises the arrangement of sensors and effectors, of joints and muscle groups, as well as of the elements and connectivity patterns of the information processing system. (2) Dynamics. Experience-independent mechanisms of growth and plasticity shape body morphology and neural connectivity, even in the absence of external environmental signals. This type of developmental change dynamically shapes morphology and allows progressively more refined mappings of stimuli and motor patterns. (3) Behavior. The mapping of sensory patterns onto motor patterns results in a primary behavioral repertoire. This repertoire allows the organism to respond properly to an evolutionarily anticipated set of environmental stimuli. In that respect, this primary behavioral repertoire reflects a coarse metric of value assignment, which establishes the organism's capacity for assigning semantic value to environmental situations, a prerequisite of the capacity for subjective evaluation and representation.

In combination, these three dimensions of innate knowledge contribute to innate value. We conceptualize value as a bias reflecting the global evaluation of recent behavior, and the capacity of a response to increase the likelihood that it will recur in the same stimulus context (Friston et al. 1994). Value, in this sense, is analogous to fitness in evolution. If we consider that neural activation patterns can be mapped onto behavioral responses, value then defines the shape of an adaptive landscape (Sporns and Edelman 1993) with peaks of high value (associated with "valuable" neuronal response patterns and their concomitant behaviors) and valleys of low value (associated with neuronal patterns that do not generate value). The shape of this value landscape is determined by innate knowledge, as well as by innate internal needs or "biases." Value systems encode value signals and their action modifies probabilities of behavior, given certain inputs. Innate value signals are elicited by rewards, or by noxious or painful stimuli. The innate ability to sense and represent value provides a major ingredient for setting up a primary repertoire of internal motivational states.

2.2 Self-Organization of Acquired Knowledge

These elements of prior or innate knowledge can provide powerful means for structuring simple behaviors and may be sufficient for adjusting the frequency of some behavioral responses according to simple innate evaluations. However, it is fundamentally impossible for all of behavioral and cognitive control in higher organisms or robots to rely on innate knowledge alone (Friston et al. 1994; Sporns et al. 2000). Additional principles enabling self-organization and autonomous growth are needed, for two main reasons. First, innate knowledge cannot generally be exhaustive and complete – much about an organism's econiche is "unknowable" in advance, and also cannot be anticipated by evolutionary processes. Thus, knowledge representations must remain flexible and dynamic, capable of incorporating specific features of individual experience and of historically unique environments. Second, no mechanism exists by which the entire knowledge structure of an organism or robot can be created in one step (for example, by design). Instead, gradual ontogenetic development is essential in building a cognitive architecture that matches the specific requirements of the adaptive system.

We suggest that two main components of self-organization need to unfold in parallel. The first involves the elaboration of subcortical and cortical networks, from simple to complex. The second involves the development of value systems to generate acquired value, in order to ensure the highest possible flexibility and diversity in the behavioral repertoire. Both of these processes must unfold in parallel, in order to ensure the continuity of knowledge representations across extended time periods, as well as the internal consistency of encoding schemes and cross-mappings within the representational architecture.

The ontogenetic development of sensory and motor hierarchies unfolds over extended periods of developmental time. Myelination studies (Fuster 2006) have demonstrated a tendency for "lower" areas in the hierarchy (primary sensory and motor) to develop earlier, immediately followed by several polymodal and limbic "higher" areas, which have strong relations to value circuits. Then successive layers of the hierarchy are added in between these top and bottom levels. This asynchrony in ontogenetic development poses constraints on the development of a hierarchically organized value system, since it requires the existence of appropriate sensory and motor representations in order to successfully create new linkages between them. Thus, autonomous development of value systems must go in parallel with the ontogenetic refinement and maturation of increasingly complex stimulus and response representations.

As discussed earlier, value systems generate signals that are used to adjust the probabilities of behaviors by modulating synaptic changes within neuronal networks through value-dependent learning. Whether or not a stimulus is valuable or salient to the organism must depend in many cases on the organism's individual experience and cannot be a fixed characteristic of the stimulus itself. Value is subject to change over the course of learning and development. Value systems cannot exclusively rely on prewired inputs from sensory regions to generate their signals, but must incorporate

activity- and experience-dependent processes. This motivates the distinction between innate and acquired value. Innate value was viewed as evolutionarily determined, similar to an innate bias. Behaviors that satisfy homeostatic or appetitive needs, consummatory activities, or avoidance of noxious stimuli predominantly reflect prior evolutionary selection and are therefore in most cases independent of learning or experience. An innate value system provides a first coarse metric for mapping prototypic sensory situations onto purposive behavior, assigning semantic value to the respective sensory situation. Utilizing this a priori semantic metric assignment, the step-by-step specification of a more refined value assignment in the course of interactive, explorative behavior does not face the symbol grounding problem. The organism is now in a position to derive a semantic evaluation by itself, based on the innate value and modified by experience gained from explorative behavior. Such innate value, however, cannot reflect the specific configuration of the environment and cannot include stimuli that are themselves initially neutral but, in a specific environmental context, become predictive of future valuable events. Acquired value is activity-dependent and allows the value system to become sensitive to stimuli that are not able, by themselves, to trigger a value-related response.

The function of the mammalian midbrain and forebrain dopamine system in reward conditioning has been studied extensively in recent years (Schultz et al. 1997; Schultz 2002) and provides a neurobiological example of how a value system may alter its response profile according to experience. One of the components of this system, the ventral tegmental area (VTA), consists of interacting populations of GABAergic and dopaminergic neurons. Dopaminergic VTA neurons project to widespread cortical areas, including prefrontal cortex, the nucleus accumbens (NAc) and the amygdala. A large subclass of VTA dopamine neurons shows phasic, burst-like activation (Kiyatkin and Rebec 1998), often in response to primary rewards (Schultz et al. 1997; Schultz 2002). Their response pattern undergoes characteristic changes during learning. Phasic activation following primary reward does not occur when the reward is reliably preceded by other reward-predicting stimuli. These "acquired" phasic responses occur at the onset of stimuli that are "predictive" of rewards. Thus, dopamine responses are "transferred in time" to conditioned stimuli and become attenuated or disappear entirely for completely predicted primary rewards. If a fully predicted reward does not occur, dopamine neurons exhibit a transient depression of their baseline discharge rate at the time of the expected occurrence of the reward. This last finding suggests that the dopamine system has access to information concerning the timing of sensory inputs relative to the occurrence of reward.

Several computational models of the midbrain dopamine system have been proposed (Schultz et al. 1997; Montague et al. 1996), forging a strong connection between dopaminergic responses and temporal difference learning (Friston et al. 1994; Sutton 1988; Sutton and Barto 1990; Sutton and Barto 1998). Essentially, these models represent dopamine activity as signaling the current prediction error of future reward. A functional model of the dopamine system has been implemented in an autonomous robot (Sporns and Alexander 2002; Alexander and Sporns 2002a) and has provided insight into the interplay between behavioral history and synaptic modification patterns (Alexander and Sporns 2002b, 2004).

An important architectural principle that emerges from this discussion is that the self-organizing process proceeds from the "top" to the "bottom" of the hierarchy. Basic behaviors, prior knowledge and value are implemented and represented by evolutionarily "old" neural structures, while evolutionarily younger and ontogenetically more refined structures are mostly involved in more advanced processing of sensory and motor signals, but are not part of the value circuit.

As outlined above, relatively simple creatures up until the evolutionary level of reptiles express fast and stereotypic action patterns in response to sensory triggers. They are, in a limited sense, autonomous. While their value metric and knowledge representations may adapt in the course of developmental stages, their basic structure is innate and genetically encoded and cannot be extended into new sensory or behavioral domains through experience, thus dramatically limiting their behavioral repertoire. To enable the flexibility of more sophisticated living organisms, it is not sufficient to rely exclusively on genetic evolutionary encoding or on a human designer, neither of which can completely define and encode the sophisticated relational structure of a complex value and knowledge representation. This structure has to be acquired in exploration of and interaction with the environment, by the individual organism. However, evolution can provide an additional structure (as part of innate knowledge) that can guide the creation of an increasingly sophisticated behavioral repertoire in concert with a similarly diversified value representation. At the earliest evolutionary level, amphibians possess a cortex, overlaying a sensory-behavioral automaton, as a means to observe sensory-behavioral matching (Körner et al. 1996, 1997, 1999) and its evaluation through feedback from the environment, and then memorize the experience as knowledge.

2.3 Active Exploration, Expectation and Surprise

Having such a means for storing experienced interactions with the environment is only a necessary but not a sufficient condition for enabling autonomous learning to get flexible behavior. For stored experience to be useful for improving and optimally shaping future behavior, the knowledge representation itself must have a consistent relational structure, whereby knowledge includes both experienced mappings from sensory situations to behavior as well as their semantic/pragmatic evaluation according to the concurrently developing value systems. A consistent relational structure of the knowledge representation during active, intended interaction with the environment cannot be generated by limiting learning exclusively to supervised ("teaching") modes. Teaching is an effective means to support and accelerate learning, but it is the capability for *autonomous* acquisition of experience which makes the difference. Limiting learning to a supervised acquisition of knowledge entails that the structure of represented knowledge (including its semantics) is provided from the outside – this is the typical scenario that gives rise to the symbol grounding problem (Harnad 1990). In contrast, to achieve flexible autonomous behavior, an organism or robot must actively explore its environment and it must have the capability to decide by itself: what is new,

what is to be learned, where it is to be learned, and how the new knowledge is to be integrated into the currently existing relational architecture (Körner 1994).

The elaboration of innate architectures by self-organization needs an important substrate for its continued development: the encounter of stimuli that violate expectations, and thus create novelty and surprise (Berlyne 1960; Oudeyer et al. 2007). To maintain the continuity of development (and prevent its getting "stuck" in local minima) requires the implementation of behavioral strategies that aim at exploring new sensory and motor configurations, with the express purpose of increasing variance. Thus, exploratory behavior has a crucial role to play in the development of new motor strategies (Angulo-Kinzler 2001; Piek 2002; Lungarella and Berthouze 2004) as well as the self-organization of knowledge representations. Exploration deliberately seeks to expand the limits of the behavioral repertoire beyond that of known stimulus/behavior couplings, by effectively linking novel sensorimotor patterns to patterns that are already incorporated in the existing cognitive architecture.

Exploration proceeds in two modes: First, an organism or robot may engage in previously learned behavioral patterns in novel environments or sensory configurations. This type of behavior may serve to diversify and enrich the already existing behavioral repertoire by increasing the variety of behaviors that are available in each context. Second, an organism or robot may spontaneously modify learned behaviors that are emitted in known environments. This type of behavior may help to optimize existing behaviors by uncovering previously unknown behavioral strategies. The fundamental role of these two modes of exploratory behavior, as generators of diversity, underscores the relationship of the self-organization of knowledge representations to systems driven by variation and selection (Sporns 1994). All such systems require a "generator of diversity" in order to maintain a repertoire of variable patterns that can be subjected to an evaluation of fitness. If variability is reduced or eliminated, selectional systems become frozen and are unable to develop further.

2.4 Self-Referential Control

Compared to simple creatures acting like reflex automatons, a system capable of flexible autonomous behavior requires a second type of innate knowledge which must be available from the beginning of ontogenetic development and learning: A memory architecture whose dynamics is strictly under self-referential control that ensures the consistency of the autonomously emerging relational structure of both knowledge and value representation (Körner and Matsumoto 2002; Körner 1994). In this context, self-reference means that, for any sensory input, it is not externally defined criteria, but the system itself that decides what is new, and if it is to be memorized, in which relation to the already acquired knowledge and where within the relational architecture the new knowledge must be integrated.

The key argument for this constraint is similar to the one outlined above for the need of a subjective value system. The relational structure of the knowledge to be stored could be defined from outside, by the designer of the system and/or the supervisor

of the learning procedure. However, imprinting a structure onto the knowledge representation which is not inherently consistent with the innate value system and the innate control architecture (algorithms) for the acquisition of knowledge will deprive the system of the capability for autonomous acquisition of knowledge. Once the relational structure of the knowledge representation is defined from outside of the system and not based on an intrinsic order, any subsequent storage of new knowledge has to be supervised from the outside as well. In contrast, the system can perform autonomous acquisition of new knowledge and integrate it into the already existing relational structure of its knowledge representation when utilizing the starting conditions of an innate value system and of innate knowledge provided in the form of a genetically encoded control architecture for enforcing a consistent subjective knowledge representation – the self-referential control architecture.

What is self-referential control about? We propose that to ensure a consistent relational structure of the subjective knowledge representation, a kind of "representational immune system" is required. In animals, the immune system controls the necessary chemical communication of the "self" with its environment, letting pass all chemical substances which have a structure compatible with the "self," while filtering out structures which are foreign and thus incompatible with the established chemical organization of the body. A representational immune system would cross-reference sensory inputs with respect to an already established representational architecture, decomposing the input into known (or compatible) parts by top-down prediction, and isolating the unknown (or incompatible) residual. The identification of predicted components of inputs serves to activate those locations where the respective knowledge elements (patterns) are stored including the activation of cross-links within the representation and to the value system, and, hence, defining what to store, where to store it, and how to integrate the new pattern (the residual) into the relational architecture of the already acquired knowledge.

3 Conclusions

Value and self-referential control constitute necessary ingredients for the autonomous development of intelligent behavior, and, therefore, for the design of humanoid robots with advanced cognitive and behavioral capacities. Our basic approach outlined in this chapter involves the identification of the principles of autonomous development that are found in biological organisms, including humans, and attempt to apply them in models of cognition. Our approach deals with questions such as: How does cognition grow as the brain builds its own connectivity and begins to interact with the surrounding world? How does the emerging cognitive architecture sample information about the saliency of behaviors and stimuli, and how does this saliency contribute to an internal motivational state? How does a neural system generate and maintain a persistent internal state that can guide behavior?

We consider three major theoretical aspects of cognition as fundamental: (a) cognition depends on knowledge representations that develop autonomously within

the complex hierarchical structure of brain networks, depending on distributed processes (see e.g., Edelman 1987; Tononi et al. 1998; Sporns et al. 2004); (b) cognition is crucially dependent on value systems, which define the saliency of sensory and behavioral representations and serve to allocate neural and bodily resources (see e.g., Friston et al. 1994; Körner and Matsumoto 2002; Sporns and Alexander 2002; Körner et al. 1996; Edelman 1987); and (c) cognition is fundamentally embodied, and cannot be disconnected from sensory-motor and bodily interactions, a natural consequence of the dynamic coupling between brain, body and world (see e.g., Chiel and Beer 1997; Sporns et al. 2000; Reeke et al. 2005).

These theoretical issues have far-reaching implications for our attempts to design intelligent systems. They demand that we abandon attempts to create artificial cognition by designing collections of isolated functional modules "running" separate computational processes that do not connect to each other, or to an organism's/robot's internal motivations, or to the environment in which the organism/robot is situated. Perceptual, cognitive and behavioral capabilities at any time within an organism's lifeline are tied to the structure of its brain and body. Importantly, the architecture of the brain, as well as the morphology of the body and the statistics of the environment, is not completely fixed. Rather, beyond a genetically determined general functional architecture the brain connectivity is subject to a broad spectrum of input-, experience-, and activity-dependent processes which shape and structure its patterning and strengths. These changes, in turn, result in altered interactions with the environment, exerting causal influences on what is experienced and sensed in the future. Value systems are crucial for guiding and shaping this process. Value gives direction to the behaving system and organizes its developmental process – while at the same time being a part of the developmental process, and being shaped by experience and individual history. As such, value systems are indispensable components of any self-organizing cognitive architecture.

In summary, both a basic configuration of a behavioral repertoire and an innate value system as described above constitute necessary innate knowledge that ensures basic capabilities of autonomous system behavior. Additionally, having a neural structure such as the cerebral cortex which is capable of self-referential control opens the possibility for autonomous learning incrementally to create a more and more complex and yet consistent subjective relational knowledge and value representation, with value being one specific quality of acquired knowledge in this process of becoming more intelligent.

References

Alexander WH, Sporns O (2002a) An embodied model of learning, plasticity and reward. Adapt Behav 10:143–159

Alexander WH, Sporns O (2002b) Timed delivery of reward signals in an autonomous robot. In: Hallam B, Floreano D, Hallam J, Hayes G, Meyer J-A (eds) Animals to animats 7: proceedings of the seventh international conference on the simulation of adaptive behavior. MIT, Cambridge, MA, pp 195–204

Alexander WH, Sporns O (2004) Interactions of environment, behavior, and synaptic patterns in a neuro-robotic model. In Schaal S, Ijspeert A, Billard A, Vijayakumar S, Hallam J, Meyer J-A (eds) From animals to animats 8, Proceedings SAB 2004. MIT, Cambridge, MA, pp 13–22
Angulo-Kinzler RM (2001) Exploration and selection of intralimb coordination patterns in 3-month-old infants. J Motor Behav 33(4):363–376
Arbib MA, Fellous J-M (2004) Emotions: from brain to robots. Trends Cogn Sci 8:554–561
Asada M, MacDorman K, Ishiguro H, Kuniyoshi Y (2001) Cognitive developmental robotics as a new paradigm for the design of humanoid robots. Robot Auton Syst 37:185–193
Berlyne DE (1960) Conflict, arousal, and curiosity. McGraw Hill, New York
Berridge KC (2004) Motivation concepts in behavioral neuroscience. Physiol Behav 81:179–209
Birmingham JT, Tauck DL (2003) Neuromodulation in invertebrate sensory systems: from biophysics to behaviour. J Exp Biol 206:3541–3546
Chiel HJ, Beer RD (1997) The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. Trends Neurosci 20:553–557
Dalgleish T (2004) The emotional brain. Nat Rev Neurosci 5:583–589
Dolan RJ (2002) Emotion, cognition, and behavior. Science 298:1191–1194
Doya K (2000) Metalearning, neuromodulation, and emotion. In Hatano G, Okada N, Tanabe H (eds) Affective minds. Elsevier, Amsterdam, pp 101–104
Edelman GM (1987) Neural darwinism. Basic Books, New York, NY
Edelman GM (1988) Topobiology. Basic Books, New York, NY
Fellous J-M, Linster C (1998) Computational models of neuromodulation. Neural Comput 10:771–805
Friston KJ, Tononi G, Reeke GN Jr, Sporns O, Edelman GM (1994) Value-dependent selection in the brain: simulation in a synthetic neural model. Neuroscience 59:229–243
Fuster JM (2006) The cognit: a network model of cortical representation. Int J Psychophysiol 60:125–132
Hammer M (1993) An identified neuron mediates the unconditioned stimulus in associative olfactory learning in honeybees. Nature 366:59–63
Hammer M (1997) The neural basis of associative reward learning in honeybees. TINS 20:245–252
Harnad S (1990) The symbol grounding problem. Physica D 42:335–346
Hasselmo ME (1995) Neuromodulation and cortical function: modeling the physiological basis of behavior. Behav Brain Res 67:1–27
Hasselmo ME, Wyble BP, Fransen E (2002) Neuromodulation in mammalian nervous systems. In: Arbib M (ed) Handbook of brain theory and neural networks, 2nd edn. MIT, Cambridge, MA
Huang X, Weng J (2002) Novelty and reinforcement learning in the value system of developmental robots. In Prince CG, Demiris Y, Marom Y, Kozima H, Balkenius C (eds) Proceedings of the second international workshop on epigenetic robotics: modeling cognitive development in robotic systems 94, Edinburgh, Scotland, pp 47–55
Kiyatkin EA, Rebec GV (1998) Heterogeneity of ventral tegmental area neurons: single-unit recording and iontophoresis in awake, unrestrained rats. Neuroscience 85:1285–1309
Körner E (1994) Autonomous recognition and selforganization of knowledge representation in neural networks. Part 1: from structuring data by setting constraints to constraint generation by self-referential control. Holonics 4:3–34
Körner E, Körner U, Matsumoto G (1996) Top-down selforganization of semantic constraints for knowledge representation in autonomous systems: a model on the role of an emotional system in brains. Bull Electrotech Lab 60:405–409
Körner E, Tsujino H, Matsutani T (1997) A cortical-type modular neural network for hypothetical reasoning. Neural Netw 10:791–814
Körner E, Gewaltig MO, Körner U, Richter A, Rodemann T (1999) A model of computation in neocortical architecture. Neural Netw 12:989–1005
Körner E, Matsumoto G (2002) Cortical architecture and self-referential control for brain-like computation. IEEE Eng Med Biol 10:121–133
LeDoux J (1996) The emotional brain. Simon and Schuster, New York

Lungarella M, Berthouze L (2004) Robot bouncing: on the synergy between neural and body-environment dynamics. In Iida F, Pfeifer R, Steels L, Kuniyoshi Y (eds) Embodied artificial intelligence. Springer, Berlin, pp 86–97

Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. Connect Sci 15:151–190

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. J Neurosci 16:1936–1947

Oudeyer P-Y, Kaplan F, Hafner VV (2007) Intrinsic motivation systems for autonomous mental development. IEEE Trans Evol Comput 11(1):265–286

Piek JP (2002) The role of variability in early development. Infant Behav Dev 156:1–14

Reeke GN Jr, Sporns O, Edelman GM (1990) Synthetic neural modeling: the "Darwin" series of recognition automata. Proc IEEE 78:1498–1530

Reeke GN, Poznanski RR, Lindsay KA, Rosenberg JR, Sporns O (2005) Modeling in the neurosciences: from biological systems to neuromimetic robotics. CRC, London

Rolls ET (1999) The brain and emotion. Oxford University Press, Oxford

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36:241–263

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599

Servan-Schreiber D, Printz H, Cohen JD (1990) A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. Science 249:892–895

Seth A, Sporns O, Krichmar J (2005) Neurorobotic models in neuroscience and neuroinformatics. Neuroinformatics 3:167–170

Sporns O (1994) Selectional and instructional ideas in neuroscience. Int Rev Neurosci 37:3–26

Sporns O (2005) Developing neuro-robotic models. In: Mareschal D (ed) Neuroconstructivism, vol 2: Perspectives and prospects. Oxford University Press, Oxford

Sporns O, Alexander WH (2002) Neuromodulation and plasticity in an autonomous robot. Neural Netw 15:761–774

Sporns O, Edelman GM (1993) Solving bernstein's problem: a proposal for the development of coordinated movement by selection. Child Dev 64:960–981

Sporns O, Almassy N, Edelman GM (2000) Plasticity in value systems and its role in adaptive behavior. Adapt Behav 8:129–148

Sporns O, Chialvo D, Kaiser M, Hilgetag CC (2004) Organization, development and function of complex brain networks. Trends Cogn Sci 8:418–425

Sugrue LP, Corrado GS, Newsome WT (2005) Choosing the greater of two goods: neural currencies for valuation and decision making. Nat Rev Neurosci 6:363–375

Sutton RS (1988) Learning to predict by the methods of temporal difference. Mach Learn 3:9–44

Sutton RS, Barto AG (1990) Time derivative models of Pavlovian reinforcement. In Gabriel M, Moore J (eds) Learning and computational neuroscience: foundations of adaptive networks. MIT, Cambridge, MA, pp 539–602

Sutton RS, Barto AG (1998) Reinforcement learning. MIT, Cambridge, MA

Tinbergen N (1951) The study of instinct. Oxford University Press, New York

Toates F (1986) Motivational systems. Cambridge University Press, Cambridge

Tononi G, Edelman GM, Sporns O (1998) Complexity and the integration of information in the brain. Trends Cogn Sci 2:44–52

Weng J, Zeng S (2005) A theory of developmental mental architecture and the Dav architecture design. Int J Hum Robot 2:145–179

Weng J, McClelland J, Pentland A, Sporns O, Stockman I, Sur M, Thelen E (2001) Autonomous mental development by robots and animals. Science 291:599–600

Cortical Connectivity: The Infrastructure of Thoughts

Giorgio M. Innocenti

1 Thinking: The Associational Nature of Thinking

This chapter rests on the following, somewhat speculative, considerations:

1. Thinking is the association of learnt and perceived sensory-motor items, with motivations and emotions, in variable proportions, into a form (Gestalt).
2. Cerebral cortex thinks. Thoughts are implemented by cortical neuronal assemblies, usually distributed over the two hemispheres.
3. Cortico–cortical connections implement the associative power of cerebral cortex, i.e., neuronal assemblies.
4. The computations performed by cortical neuronal assemblies are identical irrespective of the location of their neurons.

Some qualifications are necessary. Thinking creates "thoughts" by associating sensory and/or motor items, some of which are memorized, and others that are present in experience. Their proportion varies but they are all necessary in thinking. Thus, the solution of a mathematical, philosophical, or artistic problem associates mainly memorized items with each other and to a lesser degree with visuo-motor schemes or models. In contrast, thinking during a tennis match associates mainly perceptual information on the position of the opponent, trajectory, speed and spin of balls with motor schemes, and to a lesser degree with experience of the same or of similar opponents. Eventually, thinking relaxes into a form ("Gestalt", thought or idea), i.e., an individual entity with some degree of completeness (closure) emerging, from an unstructured, fragmented mental background. Creative thinking results from previously unexplored associations leading to novel thoughts. Motivations and their counterpart emotions can dominate, and occasionally hamper, thinking, as in the case of trying to solve conflicts. In contrast, emotions can be limited to the experience of surprise as when a new idea emerges, which was neither thought of

G.M. Innocenti
Department of Neuroscience, Karolinska Institutet, Stockholm, Sweden

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, On Thinking,
DOI 10.1007/978-3-642-03129-8_22, © Springer-Verlag Berlin Heidelberg 2010

nor of any practical use, as a key component of humor. Consciousness is often, but not always, a property of thinking.

2 Neuronal Assemblies: An Operational Definition

All brain operations result from the formation of neuronal assemblies, but higher brain functions, such as thinking, involve assemblies of cortical neurons. This does not mean that subcortical structures do not participate in thinking. Some of them do, in particular the thalamic nuclei. The thalamus is probably crucial in providing the level and pattern of activity necessary for the formation of cortical neuronal assemblies. However, the thalamic nuclei are usually embedded in "labeled lines" carrying specific sensory or motor information to the cortex. Therefore, they lack the broad span across items of different nature which characterize thinking. The experimental work described below had the goal of dissociating the thalamic from the cortico–cortical components in assembly formation. We chose to study assemblies distributed over the two hemispheres since the bihemispheric thalamic projection is limited or nonexistent. As a rule, thinking involves the two hemispheres and therefore must use the connections between them. These connections course through the corpus callosum and provide the structural basis for interaction among cortical neurons.

A neuronal assembly is a morpho-physiological entity implementing what Mountcastle (1978) called a "distributed system". It consists of discrete neuronal groups (columns or clusters) distributed over one or several areas, whose activity can become associated in transient, flexible ways. Thus, each neuronal group can become involved in different assemblies within different, often sequential, time frames. Not all neurons in an assembly necessarily participate by spiking; some might be depolarized below spiking threshold, creating a subliminally excited field out of which spike-conveyed messages are carved out. The assembly consists of excitatory and inhibitory neurons in variable proportions. What characterizes the assembly is that its constitutive neuronal groups together implement a time-discrete fraction of a perceptual, motor, or cognitive process, with some degree of completeness (closure) emerging as a unit from background cortical activity. Intentionality or consciousness are not necessary in the emergence of a neuronal assembly as shown by hallucinations, dreams, coordinated unintentional movements and sudden, unexpected thoughts. Which kind of task, perceptual, motor, cognitive, etc., the assemblies perform depends on the distribution of the constituent neurons across the different cortical areas, although the assemblies perform identical or similar computations, irrespective of their position. What we mean by computation will be clarified below. A distinction can be drawn from the onset between neuronal assemblies characterized by synchronous activity, such as those presumably underlying figure-ground discrimination in perception (Maldonado et al. 2008 and references therein) or recall, and those implementing sequential processes, as in motor functions. Our concern is with the first.

3 Cortico–Cortical Connections: Axonal Geometries; Elementary Axonal Computations; Development

Neuronal assemblies are implemented via cortico–cortical connections and to some extent by cortico–thalamo–cortical loops (Guillery 1995).

During the last 30 years, the usage of axonal transport of different molecules in experimental animals has generated an extensive description of the topography of cortico–cortical connections. This has resulted in connectional matrices or models of extraordinary complexity (see Felleman and Van Essen 1991; Scannell and Young 1993; Stephan et al. 2000). Yet, impressive as they are, those results have only scratched the tip of the iceberg. Recent techniques allowing the detailed reconstruction and quantification of individual axons from their site of origin to their termination has opened a new epoch in connectional studies. These studies have brought attention to the complexity of axonal geometries. Axons, far from performing a faithful transfer of information from the neuronal cell body to its targets, as it would be required of an electric cable, participate in the computation performed by assemblies of cortical neurons. They implement at least three types of operations: mapping, differential amplification and delaying (Innocenti 1995).

Mapping results of that a point in the cortex, i.e., the location of a cortical neuron is mapped onto the spatial distribution of the synapses in the target territory of its axonal arbor. Differential amplification results of **that** certain targets of an axon receive more, and others fewer, synapses. Delaying is due to the fact that individual axonal arbors are made of branches of different length and diameter, which, by causing differential conduction delays, affect the timing of activation of the targets. In addition, since cortical axons come in a large spectrum of diameters the potential for computations exploiting differential activation delays appears to be great, and indeed might be a specific feature of cortical design. One other kind of computation might be implemented by axonal geometry: frequency filtering at axonal bifurcations (Parnas and Segev 1979; Lüscher and Shiner 1990). In addition it remains to be clarified if all the synapses generated by a single axon share the same probability, quality, quantity, and speed of neurotransmitter release.

The detailed study of axonal arbors has led to the view that axonal arbors consist of two compartments (Tettoni et al. 1998): (1) a conduction compartment, i.e., the synapse-free branches responsible for communication within the axon, and, (2) a transmission compartment i.e., terminal or preterminal branches loaded with synaptic boutons, responsible for communication among neurons. Surprisingly, axons as different from each other as the thalamo–cortical axons to the barrel-field of the mouse and the visual callosal axons in the cat were found to be similar in most parameters of their terminal arbors. However, the transmission compartment was more modest and the conduction compartment larger in the callosal axons (Tettoni et al. 1998), which is consistent with their modulatory and associational functions.

The construction of axonal arbors is a late developmental event involving exuberant production of long axonal collateral, as well as of short axonal branches and synapses,

followed by positive selection of those which will be maintained (reviewed in Innocenti and Price 2005). Thalamic input, at least in sensory systems, has the power of validating the construction of the arbor. Competition among different axonal systems seems to play an additional role (Caminiti and Innocenti 1981; Restrepo et al. 2003).

Axonal connections in the primary visual areas embody at least two of the principles of perceptual grouping identified by Gestalt psychology: proximity and collinearity. Cortical connections are more abundant between nearby locations in cortex, corresponding to nearby locations in the visual field, and decrease with distance. Furthermore, they interconnect collinear columns of orientation specificity (Bosking et al. 1997; Schmidt et al. 1997). The design of cortical connectivity, therefore, seems to provide the structural basis for figure-ground segregation in perception.

4 Physiological Analysis of a Visual, Bihemispheric Neuronal Assembly: The Role of Axonal Geometries

The studies described below were aimed at characterizing synchronous neuronal assemblies generated by visual stimuli in the two hemispheres in terms of their location, dependence on stimulus configuration, temporal structure, and neuronal implementation. The latter addressed the role of callosal connections in assembly formation. The first set of studies used a stimulus consisting of collinear gratings, moving identically in the two hemifields, and therefore in the hemifield representations of the two hemispheres. These gratings conformed to the Gestalt grouping principles (Rock 1995) of proximity, collinearity, and common fate. Therefore they were easily bound into a unified percept. The results were contrasted with gratings of identical spatial frequency and speed of movement but orthogonally oriented in the two hemifields.

We tested if (1) the stimuli synchronized activity in the two hemispheres, (2) the synchronization depended on stimulus configuration, (3) the synchronization could be revealed by EEG analysis, (4) comparable results could be obtained in animals (ferrets) and man and (5) the synchronization was abolished by transection of the corpus callosum. Those studies used the analysis of EEG coherence in ferrets and in humans and responded in the affirmative to all the questions above (Kiper et al. 1999; Knyazeva et al. 1999). Only the collinear stimuli increased EEG synchronization and this was specific for the beta–gamma bands. A subsequent fMRI study in parallel with the EEG recordings detected an increased BOLD signal in the lingual and fusiform gyri (presumptive areas V4 and VP) linearly correlated with the increased coherence, when the collinear gratings were presented (Knyazeva et al. 2006). Therefore, increased metabolic activity, possibly related to increased neuronal depolarization appears to parallel the synchronization.

The rational for studying callosal connections was that they are widely believed to implement similar operations to intracortical and/or inter-areal connections

within each hemisphere (Innocenti 1986; Kennedy et al. 1991). Therefore they can provide a general model for the study of cortico–cortical connectivity. Furthermore, over the past 15 years we have provided a sufficiently detailed description of the geometry of callosal axons, their computational properties and development (Houzel et al. 1994; Innocenti et al. 1994; Aggoun-Zouaoui and Innocenti 1994; Aggoun-Zouaoui et al. 1996; Zufferey et al. 1999). Visual areas were chosen because of the large amount of information available on their structural and functional organization. The focus on synchronization was motivated by the hypothesis that synchronous activity may identify at least one type of cooperative neuronal assemblies (Gray et al. 1989; Eckhorn et al. 1988; Abeles 1991). Furthermore, simulation experiments suggested that the majority of callosal axons interconnecting the primary visual areas have geometries apt to synchronize activity in their terminal territories (Fig. 1) (Innocenti et al. 1994).

In a subsequent study a novel indicator of neuronal synchronization, the S estimator, was applied to EEG recordings performed in human subjects with high-density (128) electrodes (Carmeli et al. 2005). The S estimator is grounded in the dynamic system theory; it corresponds roughly to 1- Entropy. It is multivariate, therefore can be applied to clusters of electrodes, and is frequency-independent; although

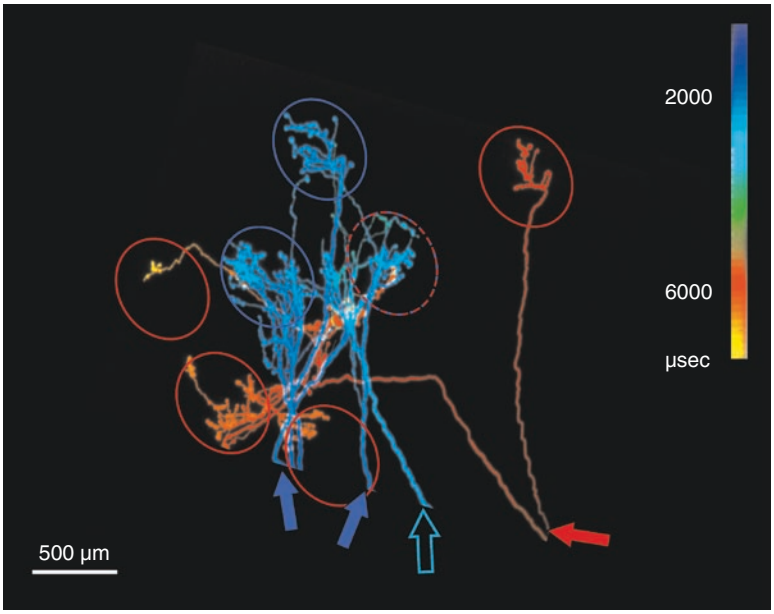

Fig. 1 Three callosal axons terminating near the border between areas 17 and 10 in the cat were reconstructed from serial sections and invaded by a simulated action potential traveling at a speed determined by length and thickness of the parent axon and of its axonal branches (Innocenti et al. 1994). Notice that each axon activates synchronously its targets (ovals are stylized orientation columns; times of activation are *color-coded*) but two of the axons (denoted by *red* and *blue arrows*) activate their targets with a delay of about 4 ms

the different frequency components of the EEG can also be separately assessed. The S estimator showed that the collinear gratings cause an interhemipheric cluster of synchronized activity in the occipital electrodes, in the beta and gamma bands. The orthogonal gratings caused synchronous activity in parieto–temporal electrodes in the alpha band and in the occipital electrodes in the gamma band.

In further experiments in ferrets (Carmeli et al. 2007) we tested the role of interhemispheric connections in controlling the local stimulus-induced synchronization in the contralateral hemisphere. We recorded field potentials near the border between areas 17 and 18, which is strongly callosally connected (Manger et al. 2002) using an array of 15 microelectrodes. Collinear or orthogonal gratings were presented to the two hemifields. The S estimator showed that the stimulus desynchronized the local activity compared to a blank screen. The magnitude of stimulus-induced desynchronization was modulated by input from the other hemisphere, but only during the presentation of collinear gratings. This was shown by inactivating the visual areas contralateral to the recording site by reversible cooling. In about 50% of the cases stimulus-induced desynchronization was increased by cooling the contralateral areas, suggesting that we had eliminated a synchronizing input from the contralateral hemisphere. This was in line with the geometry of individual callosal axons (above). However, in 50% of the cases inactivation of the contralateral areas had a synchronizing effect. In this case therefore the input from the contralateral cortex had a desynchronizing effect. The most likely explanation of this dual effect was that the electrode was sampling activity elicited by two callosal axons with different conduction velocities (Fig. 2).

In a third study (Makarov et al. 2007) we analyzed the consequences of inactivating the contralateral visual areas on the local field potentials recorded at the individual electrodes of the multielectrode array. The effects were again stimulus-dependent and consisted of a combination of enhancement and depression of the responses, the first predominating at short latency after the stimulus onset and the second at longer latency. Thus excitatory–inhibitory interactions between the hemispheres appear to underlie the formation of synchronous neuronal assemblies.

5 Implications of Principles of Brain Organization for Thinking: Cortico–Cortical Connections May Constrain and Channel Both Perception and Thinking

The relevance of the results described above for understanding the neural implementation of thinking rests on the proposal that the computations performed by neuronal assemblies are identical irrespective of the location of their neurons in the various subdivision of cerebral cortex. Computation here means: the dynamic operations the assemblies perform in spatio–temporal cortical domain.

Our findings suggest that cortico–cortical connections implement the synchronous neuronal assemblies which enact figure–ground segregation. Each assembly

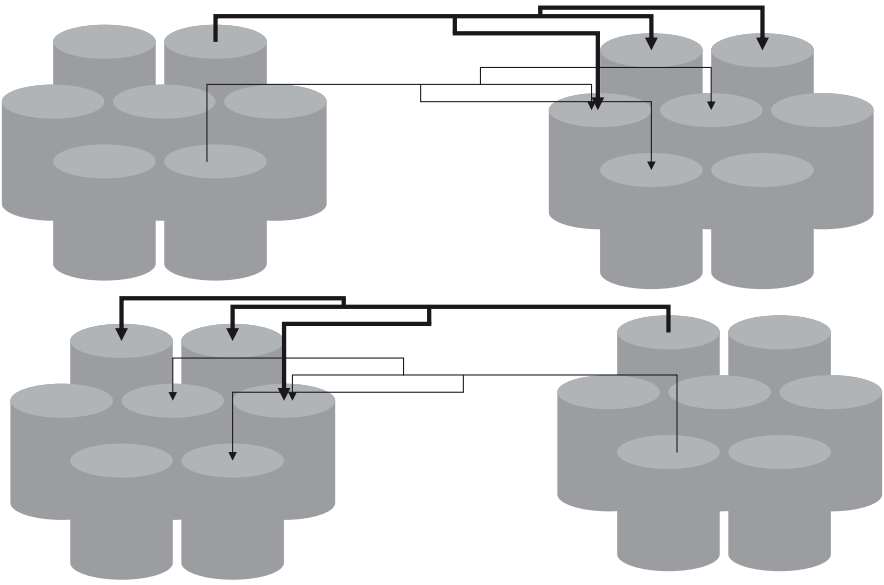

Fig. 2 Schematic sets of orientation columns in the two hemispheres are reciprocally connected with axons of different thickness, and hence conduction velocities. As in Fig. 1, each axon activates its target columns synchronously but there is a delay in the activation elicited by the two axons. It is assumed that when each axon is active separately, the net result is synchronization,, while when the two axonal systems are both active, they desynchronize their targets (Carmeli et al. 2007)

has a specific spatio–temporal dynamic signature, determined by the computational properties of the axons i.e., the selective connectivity, conduction velocity, and geometry of axonal arbors.

Therefore, cortico–cortical axons constrain and channel visual function in primary visual areas by generating perceptual building blocks in the absence of which most visual operations would be impossible. And we can postulate that thinking emerges from the formation of neuronal assemblies, by mechanisms similar to those described for perceptual neuronal assemblies, but associating sensory and/or motor items, some of which memorized, others present in experience with motivations and emotions, in variable proportions, into a form.

Thus, in thinking, cortico–cortical connections perform similar associative operations as in perception. The implications of this hypothesis are that thinking is the projection onto the world of cortico–cortical connectivity rather as perception is the projection onto the world of cortico–cortical visual connections. Cortico–cortical connectivity, however, must generate percepts or thoughts compatible with the "real world". Thus, cortico–cortical connections are under a double selective screening. The first is performed by evolution and second by development (for development see Innocenti and Price 2005).

The hypotheses above suggest that, the projections into the "real world" of some aspect of cortical organization might have left traces in the early history of mankind. Indeed, evidence that at least for visual perception this may be the case, can be found in the many nonfigurative examples of Paleolithic art some of which may be the result of endogenous cortical activities, akin to those generated by migraine or hallucinatory drugs, projecting into the world, aspects of the morpho-functional organization and connectivity of primary visual areas (Lewis-Williams 2002).

On the other hand, if perception and thinking are constrained by the structure of cortico–cortical connections, which in turn were selected in evolution, and development from a wider set (above), this amounts to postulating the existence of unperceivable percepts and of unthinkable thoughts. Both notions are challenging. The existence of unperceivable percepts might be exemplified by the bistable figures and, better, by Escher's impossible figures. But, what are "unthinkable thoughts"? I believe that at the frontier between geniality and madness they destroy the familiar, peaceful landscape of common logics. One example is reported in the preface of Foucault's "Les mots et les choses" (Foucault 1967). It refers to a classificatory scheme in a Chinese encyclopedia attributed to Borges. Animals, says Foucault, are divided into the following classes: (a) belonging to the emperor, (b) embalmed, (c) domesticated, (d) piglets, (e) sirens, (f) fabulous, (g) free dogs, (h) included in the following classification, (i) which agitate crazily, (j) numerous, (k) drawn with a very fine camel brush, (l) etc., (m) which make love, (n) which from far look like flies (my translation from French). Another example is sentences such as "I was thinking out of my father's hair", the hallmark of the schizophrenic thinking. And, interestingly, altered cortico–cortical connectivity, including callosal connections is probably at the root of the schizophrenic condition (Innocenti et al. 2003; Mitelman et al. 2007, and references therein).

References

Abeles M (1991) Corticonics neural circuits of the cerebral cortex. Cambridge University Press, Cambridge

Aggoun-Zouaoui D, Innocenti GM (1994) Juvenile visual callosal axons in kittens display origin- and fate-related morphology and distribution of arbors. Eur J Neurosci 6:1846–1863

Aggoun-Zouaoui D, Kiper DC, Innocenti GM (1996) Growth of callosal terminal arbors in primary visual areas of the cat. Eur J Neurosci 8:1132–1148

Bosking WH, Zhang Y, Schofield B, et al (1997) Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. J Neurosci 17:2112–2127

Caminiti R, Innocenti GM (1981) The postnatal development of somatosensory callosal connections after parietal lesions of somatosensory areas. Exp Brain Res 42:53–62

Carmeli C, Knyazeva MG, Innocenti GM et al (2005) Assessment of EEG synchronization based on state-space analysis. Neuroimage 25:339–354

Carmeli C, Lopez-Aguado L, Schmidt KE et al (2007) A novel interhemispheric interaction: modulation of neuronal cooperativity in the visual areas. PLoS ONE 2:e1287

Eckhorn R, Bauer R, Jordan W et al (1988) Coherent oscillations: a mechanism for feature linking in the visual cortex? Biol Cybern 60:121–130

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47

Foucault M (1967) Le mots et le choses (Italian translation). Rizzoli, Milano

Gray CM, König P, Engel AK et al (1989) Oscillatory responses in cat visual cortex exhibit intercolumnar synchronization which reflects global stimulus properties. Nature 338:334–337

Guillery RW (1995) Anatomical evidence concerning the role of the thalamus in corticocortical communication: a brief review. J Anat 187:583–592

Houzel J-C, Milleret C, Innocenti GM (1994) Morphology of callosal axons interconnecting areas 17 and 18 of the cat. Eur J Neurosci 6:898–917

Innocenti GM (1986) General organization of callosal connections in the cerebral cortex. In: Jones EG, Peters A (eds) Cerebral cortex, vol 5. Plenum, New York

Innocenti GM (1995) Exuberant development of connections, and its possible permissive role in cortical evolution. TINS 18:397–402

Innocenti GM, Price DJ (2005) Exuberance in the development of cortical networks. Nat Neurosci Rev 6:955–965

Innocenti GM, Lehmann P, Houzel J-C (1994) Computational structure of visual callosal axons. Eur J Neurosci 6:918–935

Innocenti GM, Ansermet F, Parnas J (2003) Schizophrenia, neurodevelopment and Corpus Callosum. Mol Psychiatry 8:261–274

Kennedy H, Meissirel C, Dehay C (1991) Callosal pathways and their compliancy to general rules governing the organization of corticocortical connectivity. In: Dreher B, Robinson S (eds) Vision and visual dysfunction, vol 3: Neuroanatomy of the visual pathways and their development. Macmillan, London

Kiper DC, Knyazeva MG, Tettoni L et al (1999) Visual stimulus-dependent changes in interhemispheric EEG coherence in ferrets. J Neurophysiol 82:3082–3094

Knyazeva MG, Kiper DC, Vildavsky VJ et al (1999) Visual stimulus-dependent changes in interhemispheric EEG coherence in humans. J Neurophysiol 82:3095–3107

Knyazeva MG, Fornari E, Meuli R et al (2006) Imaging of a synchronous neuronal assembly in the human visual brain. Neuroimage 29:593–604

Lewis-Williams D (2002) The mind in the cave. Thames and Hudson, London

Lüscher HR, Shiner JS (1990) Simulation of action potential propagation in complex terminal arborizations. Biophys J 58:1389–1399

Makarov VA, Schmidt KE, Castellanos NP, et al. (2007) Stimulus-dependent interaction between the Visual Areas 17 and 18 of the two hemispheres of the Ferret (Mustela putorius). Cereb Cortex 18:1951–1960

Maldonado P, Babul C, Singer W et al (2008) Synchronization of neuronal responses in primary visual cortex of monkeys viewing natural images. J Neurophysiol 100:1523–1532

Manger PR, Kiper D, Masiello I et al (2002) The representation of the visual field in three extrastriate areas of the ferret (Mustela putorius) and the relationship of retinotopy and field boundaries to callosal connectivity. Cereb Cortex 12:423–437

Mitelman SA, Torosjan Y, Newmark RA et al (2007) Internal capsule, corpus callosum and long associative fibers in good and poor outcome schizophrenia: a diffusion tensor imaging survey. Schizophrenia Res 92:211–224

Mountcastle VB (1978) An organizing principle for cerebral function: the unit module and the distributed system. In: Edelman GM, Mountcastle VB (eds) The mindful brain. MIT, Cambridge and London

Parnas I, Segev I (1979) A mathematical model for conduction of action potentials along bifurcating axons. J Physiol 295:323–343

Restrepo CE, Manger PR, Spenger C et al (2003) Immature cortex lesions alter retinotopic maps and interhemispheric connections. Ann Neurol 54:51–65

Rock I (1995) Perception. Scientific American Library, Freeman, New York

Scannell JW, Young MP (1993) The connectional organization of neural systems in the cat cerebral cortex. Curr Biol 3:191–200

Schmidt KE, Goebel R, Löwel S et al (1997) The perceptual grouping criterion of colinearity is reflected by anisotropies of connections in the primary visual cortex. Eur J Neurosci 9:1083–1089

Stephan KE, Hilgetag C, Burns GAPC et al (2000) Computational analysis of functional connectivity between areas of primate cerebral cortex. Philos Trans R Soc Lond B 355:111–126

Tettoni L, Gheorghita-Baechler F, Bressoud R et al (1998) Constant and variable of axonal phenotype in cerebral cortex. Cereb Cortex 8:543–552

Zufferey PD, Jin F, Nakamura H et al (1999) The role of pattern vision in the development of cortico-cortical connections. Eur J Neurosci 11:2669–2688

Models as Tools to Aid Thinking

Helge Ritter

Abstract We offer a brief synoptical discussion of some major modeling methodologies and discuss these from a number of complementary dimensions of modeling and thinking: how can models aid our thinking, what are their different characteristics, how is modeling affected by the selection of features, how far is the reach of mappings and what in addition can be delivered by dynamical systems, how to cope with uncertainty, what are mechanisms of optimal inference, and how is modeling connected with learning? We conclude with a brief discussion of some inherent limitations of any model.

1 Introduction

A famous thinking puzzle goes as follows: you are given two ropes and a lighter. Each rope takes exactly 1 h to burn if lighted at one end. However, the burning speed along each rope is not guaranteed to be uniform and is not guaranteed to be the same as along the other rope – only the total burning times are equal and exactly 1 h for each rope. Given this as the only equipment, how would you measure exactly 45 min?

When we solve puzzles like this, we hardly ever physically construct the underlying situation. Instead, we create in our mind a picture that captures those elements of the problem that we deem essential for its solution. It is then this "model" that is our tool for aiding our thinking towards a solution.

This is a powerful approach that offers many benefits: in the first place, availability of a model decouples us from the costs entailed with setting up the real situation – usually, we do not have the appropriate ropes at hand. Actions on the model can be much quicker and cheaper than on real objects. And while we could not reverse the burning of a rope, we easily can do so in our mental model when searching for a solution. Finally, by omitting irrelevant detail, a model can be much more focused.

H. Ritter
CITEC – Cognitive Interaction Technology and Faculty of Technology,
Bielefeld University, Bielefeld, Germany

B.M. Glatzeder et al. (eds.), *Towards a Theory of Thinking*, *On Thinking*,
DOI 10.1007/978-3-642-03129-8_23, © Springer-Verlag Berlin Heidelberg 2010

This is not only useful for aiding our thinking, but in addition it often facilitates communication, and thereby thinking, in groups.

In the following sections, we will discuss a number of issues connected with models and their use for thinking. Specifically, we will look more closely at the following questions:

- What is a *model*, how can it be a *tool*?
- How can *models help us*?
- What *input* is needed from our side?
- What *outputs* may we expect?
- What *methods* are available out there?
- What are *known limitations*?

There are many interesting perspectives from which these questions can be viewed. Our primary perspective will be to give a sketch of what might be termed as "the art of modeling," that is, the considerations, methods and requirements that go into the process of model building for the sake of augmenting or supporting our thinking.

This may but need not involve the support by computers – while complex models will often need the use of computer simulation for their full exploitation, there are also many models that are tremendously helpful even without such means. One major category comprises mathematical models that permit some degree of mathematical analysis which can then enable insights beyond what any particular simulation might offer.

A second intriguing perspective arises from cognitive and brain sciences: to what extent do models of the environment and of situations exist in our brain and are involved in our perception and our thinking, and what are their manifestations in measurable brain activity patterns? Two major facets of this question are: what happens in our brain when it abstracts situations into simplified conscious models (such as during solving the ropes puzzle)? And is there in addition evidence for models (or "representations") of objects and events that shape our mental processes below the level of what is consciously accessible to us?

Finally, a third perspective bridges the previous two: what models can we make to describe brain activity and thinking themselves, ultimately including the processes of model making and exploitation (which requires a "modeling of modeling") themselves.

Fascinating as these questions may be, lack of space (together with lack of knowledge) will restrict us largely to focus on the first perspective, with occasional points of contact to perspectives two and three, e.g., when considering "biology-inspired" approaches to modeling, such as artificial neural networks or evolutionary algorithms.

2 Models and Thinking Economy: What Models Can Do for Us

Thinking is for mastering complexity. Thinking came only late in evolution. Compared to perception and motor abilities, our thinking abilities appear weak: while we can combine millions of color pixels into the vivid percept of a rose within the

fraction of a second, thinking becomes very hard, if not impossible, when we have to keep track of more than a handful of items and relationships simultaneously.

A large share of responsibility for this limitation lies in our working memory, which can only accommodate Miller's famous "7 plus minus one" items (Miller 1956). For a long time, it has been thought that this capacity limit is inborn and not trainable. Recent findings indicate that certain forms of training may indeed allow to extend working memory to some extent and thereby enhance thinking abilities (improve what is called "fluid intelligence") (Jaeggi et al. 2008).

However, even in the presence of such encouraging findings, working memory remains a scarce resource. Therefore, in order to make thinking efficient, it is important to fit its "objects" into its limited arena. It has been speculated that this limitation may have provided a major driving force for the development of our ability of abstraction, i.e., mapping the detailed image of reality into simplified situations in which only a manageable number of relationships is present.

The formation of a good model is akin to the process of abstraction: it provides a more "condensed" representation of what was originally given. Usually, this is accompanied by a strong degree of "data compression." For instance, with Kepler's model of planetary motion all the observation points of a planet could be compressed to a handful of parameters describing its orbit around the sun (taken the orbital parameters of the earth as known). In the history of science, the discovery of such models has marked milestones of scientific developments and has tremendously aided our thinking about natural phenomena. Finding such models was usually the result of human ingeniousness together with long and laborsome analyses of large amounts of data. More recently, the young discipline of machine learning (Bishop 2006; Witten and Frank 2005) has provided us with tools that greatly allow reducing the need of human intervention to create nontrivial, condensed models from raw data. In the following sections, we will discuss some of the major pertinent approaches, their prospects as well as their limitations.

In view of these remarks, creating a good model usually confronts us with the following questions: how can we describe our observations in a way that is

- Somehow *deeper* than data alone
- *Reflective* of the essential factors
- Sufficiently *"condensed"* to aid our thinking
- At the *right level of complexity*
- *Mathematically specific* as opposed to language
- Easy to *communicate* to others

The process of constructing models that meet these criteria can challenge our thinking in the direction of explicitness and clarity. The results can reward our thinking with complexity reduction and predictive power and augment our thinking by adding quantitative precision and making computer assistance feasible. Finally, good models help to connect our thinking by reducing ambiguity and making insights more portable between people.

3 Major Dimensions of Modeling

Even as models try to provide us with simplifications of what happens in the world around us, their abstractions still reflect a great deal of the enormous richness of our world. As a result, models can be of very different kinds. But they all share the property that they answer a certain range of questions. The type of questions that can be addressed can give us some key dimensions of modeling:

- *Structural versus functional.* Structural models answer how something is composed of simpler constituents and how these interact. We encounter such models in many disciplines, e.g., when describing the atom arrangement of a molecule, or the inner structure of some organism. Functional models go beyond depiction and try to "attach" to their constituents functionality and purpose, for instance, conceptualizing the heart not only as a hollow muscle but as a pump with a specific purpose within the organism.
- *Coarse-grained versus fine-grained.* A coarse-grained model can be very valuable to provide a good picture of the essentials that can be efficiently manipulated. When it comes to accuracy, one often needs more fine-grained models, encompassing more detail. However, if a fine-grained model represents more detail than warranted by data, it falls victim to "overfitting" and changes from a description into a "phantasy."
- *Quantitative versus qualitative.* Quantitative models endow us with the precision of numerical values. Often, this comes at the price of sacrificing pictures or functional aspects plus the need to bring questions and answers into a numerical format. In exchange for that, quantitative models allow to utilize a wide range of mathematical methods which can be a powerful complement to our intuitive thinking.
- *Descriptive versus predictive models.* While descriptive models "just" attempt to provide us with a picture of the present, predictive models attempt to extrapolate this picture into the future. Their complexity can range from predicting a single parameter (e.g., a stock value) to a large set of values (e.g., the temperature distribution across all cities in a country).
- *Deterministic versus probabilistic models.* Deterministic models resemble "clockwork-like" mechanisms. When applicable, they can offer us a highly detailed picture of a phenomenon (e.g., the motion of the stars in the sky), but many important real-world phenomena can never be described or observed accurately enough to admit such "perfect" predictions. Probabilistic models account for that limitation and offer the necessary methodologies to deal with such situations as quantitatively as possible.
- *Explicit versus implicit.* Explicit models try to expose essential mechanisms in a way that is informative to their user and that supports reasoning. Typical examples are, e.g., rule-based systems. Implicit models are more modest: they capture relationships in the style of a "black box," making successful predictions, but in a way that not necessarily aids human comprehension. Typical examples are, e.g., neural networks.
- *Continuous versus discrete.* Models can be based on continuous quantities or on discrete symbols. Often, the choice of continuous versus discrete is not dictated by

the phenomenon per se, but can be a matter of the desired "resolution": the traffic in a city may use a discrete description based on individual cars, or a "smoothed" picture modeling traffic as a continuous "flow." The preference of one type over the other may help to emphasize certain aspects (e.g., "flow"), or open up particular methods (often, discretization leads to more efficient computer algorithms).

The last distinction already illustrated that the choice of a particular model type usually influences strongly the methods that we have at our disposal for analyzing, simulating, and refining the model. This leads to a number of further modeling dimensions which are motivated by the methods intended to apply during the modeling process.

With regard to studying thinking, it is very likely that we will have to draw on all these model types. In fact, each model type above can be seen as the reflection of a particular thinking mode. Therefore, the above taxonomy of model types can also be seen as a (highly incomplete) taxonomy of different thinking modes. Likewise, from a pragmatic "implementational perspective" when trying to replicate a comprehensive subset of thinking abilities on computers, we will have to implement an ability to use all of the above models and switch between them as part of realizing the amazing flexibility of thinking operations. This will require casting the knowledge of an experienced modeler into a suitable computational form to automate the necessary decisions which model representation to use as a function of context.

4 Modeling and Feature Selection

It can be a highly nontrivial task to decide which data do actually "capture" the relevant information that can serve as the input to a model. Usually, this step is termed feature selection (Special Issue on Variable and Feature Selection. Journal of Machine Learning Research 3 2003). The analog question for the output side can be nontrivial, too; however, often it is easier to say what one is interested to receive as an output than to say what is required to produce it.

In the context of thinking, feature selection appears as the problem of distinguishing the relevant aspects of a problem from additional information that is not useful for a solution. On a short time scale, this is connected with attention: selecting the items that should go into short term memory. Current attention models are usually formulated rather close to the signal level, i.e., at a rather low level of abstraction. It is largely an open question how to extrapolate the rather low-level feature selection methods that are at our disposal presently to the much higher levels of abstraction that are prominent in thinking.

As a result, feature selection often has elements of a trial-and-error procedure. One major aspect is to decide "how many" features are required. For efficiency reasons, it is usually desirable to work with as few features as possible; on the other hand, if little is known one may have to start with a generously large feature set and then apply methods that try to identify irrelevant feature dimensions.

One major guiding idea is to rank features according to their variation in a representative set of example cases and to keep only those whose statistical variance is largest. An important refinement is in addition to require the absence of correlations (since correlated features contain redundant information). If it is sufficient to guard against linear correlations, principal component analysis (PCA) is a well-established method to select useful features (Jolliffe 2002): it constructs a set of "principal coordinates" along which the variation of the data is uncorrelated and maximal. Each axis direction "captures" a proportion of the entire data variation and selecting the minimal subset of axes that together capture a given proportion (say 95%) of the total data variation often is a suitable recipe for feature selection.

One major limitation of PCA is that it can only construct pairwise orthogonal feature axes. ICA (independent component analysis, (Hyvärinen et al. 2001)) is a (computationally more expensive and intricate) generalization when statistical independence can only be achieved with nonorthogonal axis choices.

Both approaches share the limitation that they can provide only features that are linear combinations of the original inputs. Often essential information is better captured in nonlinear combinations. Self-organizing Feature maps (SOMs, (Kohonen 2001; Ritter et al. 1992) essential features also from data distributions with a strongly nonlinear geometry. While the SOM has been inspired by models of adaptivity in neural layers, more recent approaches (Roweis and Saul 2000; Meinicke et al. 2005) feature extraction take their motivation from statistical machine learning theory or are based on information theoretic measures, such as mutual entropy with respect to the desired output (Duch et al. 2004).

4.1 The Power of Mappings

Once "questions" and "answers" have the form of variables $x \in X$ and $y \in Y$ from suitably chosen feature spaces, constructing a model can be thought of as constructing a "mapping"

$$f : x \to y = f(x) \tag{1}$$

that maps input questions to their answers.

While there are situations for which a mapping will not suffice (cf. below), this formalization has turned out to be very fruitful and powerful for a large range of important modeling situations.

From the perspective of thinking, a major question of interest is to what extent thinking operations can be conceptualized in terms of such mapping models. This question can be considered at a number of levels.

One major level is to consider decisions as mappings: in this case, the input comprises the features deemed relevant for the decision, while the output usually is from a discrete set of alternatives. Closely related is the task of classification and/or categorization: here the decision is the assignment of some class or category membership of an input.

While decision-making, classification and categorization certainly form only a rather restricted subset of our thinking operations, their modeling by way of a mapping function offers a firm basis for a closer analysis. One line of analysis has been concerned with the question of optimality: what are suitable criteria for mappings that characterize optimal decisions or classifications? A second line of analysis has been along the structure of the mapping function: are there particular "families" of functions that offer good models to the way we ourselves perform decisions and make classifications during thinking? And finally, a third line of analysis has been concerned with the actual generation of mapping functions when a decision or classification problem is given. This has turned out to be closely connected with the issue of learning such mappings from known "training examples," since it is usually only such "implicit" information that is at hand when asking for a decision or classification function.

More precisely, one usually knows a certain "example set" T of associated (x, y) pairs. Finding a model then boils down to the determination of a function that is "compatible" with this example set and that makes good predictions for novel questions x not seen in the example set.

Here, "compatible" usually means a weaker condition than $f(x) = y$ for the given (x, y) pairs, since real-world data never can be perfectly accurate.

A frequently adopted criterion is to minimize the average square distance $(f(x) - y)^2$ (Bishop 2006). Here, (1) the average is taken over the example set T, and (2) the minimization is with respect to a set F of "admissible" functions f.

From a mathematical point of view, this is an optimization problem in a "function space." Practically, such optimization problems are often solved in a step-wise, iterative fashion, starting with an initial function f which is gradually "refined." This refinement can then be viewed as a form of learning: starting with a "coarse model" (initial f), subsequent refinement steps lead to models that are increasingly well in accordance with the example set.

Consequently, this form of the problem has become a key focus of *Machine Learning* (Ripley 1996). Many important questions are connected with the choices of the example set T and the set F of admissible functions:

- The "richer" F, the more models are available. On the one hand, this makes it more likely that the "correct" model (or a good approximation to it) is contained in F. On the other hand, identifying this "best" model among a larger set of competitors requires more information (larger example set T) than in the case of a small F.
- The example set T must provide the information to discriminate the "correct" model f. If T is too small (or F too large), the discrimination power may be too small and we may end up with many different models that all fit the data well. This is the situation of "overfitting": if we pick *some* winner in this unclear situation, it may later, when more data become available, turn out to be a "poor fit," i.e., to generalize poorly to "novel" data.

Another potential connection with thinking operations has become encouraged through connectionist models. They represent mappings as networks of many simple,

nonlinear mappings that themselves are considered as a model of the transformation effected by a single neuron. This leads to a very "distributed" representation of the overall mapping. Again, learning algorithms have been used to construct such networks from data for a variety of tasks, including (simple) inferences or representing rules of grammar for word flexion. A major insight from this research was that "rule-like" behavior is not necessarily "tied" to similarly "rule-like" representations but can as well be generated from representations that are distributed over many fine-grained constituents that do not readily reflect the structure that becomes manifested in the overall mapping. This has led to great caution in taking introspectively perceived "high-level" thinking operations as credible candidates for the underlying basis blocks and opened the scope for deeper representations that may be inaccessible to our introspection and in a format possibly not well matched to what we can easily communicate.

4.2 Linear Versus Nonlinear Models

The best-understood mappings are characterized by the property of linearity: once their answers are known for a number of input "patterns," they are automatically known for any linear superposition of these patterns (e.g., the answer to the sum of two inputs is always the sum of the answers to each input alone).

Therefore, the linear models produce their "answer" always as a linear combination of their input variables, formally

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n . \tag{2}$$

Here, model building amounts to finding a set of weighting coefficients a_i, $i = 0...n$, for which (2) provides a "good fit" to the data. Usually, a "good fit" is sought by minimizing the square of the deviations, requiring the solution of a linear system of equations, which is an easily solvable standard task which always leads to a unique and globally optimal solution. Closely connected is the use of linear equations to specify class boundaries in the form of hyperplanes. In its basic form, this requires the classes to be "linearly separable" (a separating hyperplane must exist). Minski and Papert (Minsky and Papert 1988) where among the first to analyze the consequences of this requirement (together with further properties) of linear models for their ability to model aspects of cognition, such as solving pattern discrimination tasks.

Regarding thinking, linear models are closely associated with two major thought patterns: superposition and independence. Superposition is the property that the response of a linear system to a linear superposition of its inputs always is the same linear superposition of its outputs to individual inputs alone. This is the simplest possible pattern of "generalization." As a consequence, the behavior of a linear system can be fully predicted from a small set of "basis examples" (e.g., the responses $a_i x_i$ to the inputs x_i alone). This is closely connected with independence: the superposition property means that the individual input–output behaviors do not interact with each other except for linear summation: each input can be attributed a

contribution to the output that is entirely independent from all other inputs. In this way, linear models are the simplest possible formalization of a situation when a larger problem (in this case the characterization of some input–output behavior when there are many inputs) can be split into smaller parts (characterizing the responses to each input alone) that can be solved independently and then combined (in this case by linear superposition) to obtain the solution of the original problem.

There is also a different notion of linearity in connection with thinking: "linear thinking" (as compared to "systems thinking," cf. below). This refers to envisaging a process as constituted by a strictly linear chain of causes and effects, neglecting possibilities such as branches or feedback loops. In this case, linearity refers to the topological structure of a "process graph" and has nothing in common with the notion of linearity (referring to some input–output characteristics) discussed in the present section.

Despite their limitations, linear models have found a wide range of applicability (Harrel 2001). The reason is that even nonlinear phenomena often can be approximated as linear behavior if the variation of their input variables is kept sufficiently small.

In addition, nonlinear phenomena often can be "linearized," e.g., by finding suitable nonlinear functions $\varphi_i(x_1..x_n)$ from which the nonlinear relationship can be obtained as a linear superposition:

$$y = a_0 + a_1\phi_1\left(x_1 \ldots x_n\right) + a_2\phi_2\left(x_1, \ldots, x_n\right) \ldots + a_k\phi_k\left(x_1 \ldots x_n\right). \tag{3}$$

This offers a lot of flexibility without increasing the computational complexity of the modeling task. Usually, the functions $\boldsymbol{\varphi}_k$ have to be chosen in a domain-specific manner (e.g., certain polynomials, or trigonometric functions if periodicity plays a role). Ideally, each $\boldsymbol{\varphi}_i()$ alone models already approximately some "component" of the phenomenon under consideration, and (3) then represents a "weighted mixture" of these components.

However, many phenomena are more than just the sum of their parts. Correspondingly, they require inherently nonlinear models. Usually the determination of such models from data is computationally much more involved than in the case of the linear models above (Gallant 1986). This is related to the so-called "local minima" problem: for each set of models there is an associated "error function," depicting the mismatch between the model and the data as a function of the model parameters.

For linear models, this error function is rather simple: it has a single, parabolic minimum that can be found in many computationally relatively inexpensive ways.

This desirable property ceases to hold for the majority of nonlinear models. Their error function usually resembles a complex "landscape" with many peaks and valleys. For such landscapes there is no general strategy to locate the lowest valley point. Most methods are some variation of the idea to change model parameters repeatedly along directions of "steepest descent" in the modeling error landscape. However, when the landscape offers many local minima, these methods are bound to get stuck in the "next best" local minimum, no matter how close this is to the globally best one.

For instance, this is a problem shared by many neural network models (Bishop 2006; Ripley 1996) Three-layer perceptrons, have an input–output relationship of the form

$$y = a_0 + a_1\phi_1\left(x, w_1\right) + a_2\phi_2\left(x, w_2\right) + \cdots + a_k\phi_k\left(x, w_n\right). \tag{4}$$

This resembles (3), but with the difference that now the functions $\varphi()$ also contain model parameters (namely the w_k, which then have the interpretation of "synaptic weights" for "neurons" k who form a weighted "superposition" of their "responses" $\boldsymbol{\varphi}_k(x, w_k)$ on inputs x).

Due to the above difficulties, optimization of such nonlinear mixture models has to be repeated with several choices for the starting values of the model parameters and the best found solution is kept.

However, not too long ago it has been observed that under particular conditions on the "kernel functions" $\boldsymbol{\varphi}$ (each $\boldsymbol{\varphi}(x, w)$ must be representable as a scalar product $\xi(x)\cdot\xi(w)$ with suitable mapping functions into a (usually high-dimensional) subspace of an infinite dimensional Hilbert space) the resulting nonlinear model again has an error function of a simple shape with a single minimum, albeit in a high dimensional parameter space. In addition, the location of that minimum can be found by efficient numeric procedures (so called "quadratic optimization"). Due to their high efficiency, the resulting class of nonlinear models are termed "kernel machines" (also "support vector nets") and constitute one major class of nonlinear models for which well-controllable numeric methods exist and which have become rather popular during the last decade (Schölkopf and Smola 2002; Shawe-Taylor and Cristianini 2000).

5 Beyond Mappings: Dynamical Systems

Despite their wide-ranging applicability, mapping models are limited in one major respect: they depict a static situation only and do not address the important phenomenon of *history-dependent* change. Loosely speaking, additional capability of dynamical systems (Kaplan and Glass 1995; Allgood, Sayer and Yorke 2000) time evolution.

To this end, dynamical systems introduce the concept of a "state" which summarizes the essential information about the system at a given instant. This means that the state acts as the "memory" that summarizes the impact of all previous inputs on the subsequent development of the output. Therefore, knowing that the state allows to "forget" everything that has happened before, everything of the past that is useful for predicting the future has been summarized in that state.

Therefore, the "state" becomes the main modeling abstraction of what has to be tracked for describing the time evolution of a system. Selecting the proper variables that are needed to represent the state of a system resembles the feature selection problem that was associated with the mapping models. For instance, since Newton we know that in addition to the position of a body of mass we also need its momentum (velocity times mass) in order to predict its course. And we know that we do

not need more: with regard to modeling the motion of the body, anything that has happened to the body before is completely summarized in the variable pair of (*position, momentum*). Therefore this pair can serve as the state of the body when we wish to model its course.

It is now well understood that purposeful body movements require the brain to control an exceedingly complex dynamical system whose state dimensionality counts in the hundreds. Therefore, on the one hand our brain is extremely well prepared to cope with dynamical systems of rather high dimensionality. However, this capability seems connected with a high degree of specialization on our movement apparatus (and some other dynamical systems associated, e.g., with the interpretation of pixel images or metabolic processes in the body).

A similar capacity seems to be largely unavailable to our conscious thinking processes. Successfully conceptualizing dynamical systems behavior and coping with it is viewed as a major prerequisite for "systems thinking," i.e., the ability to go beyond "linear thinking" and take effects such as feedback loops and memory effects into account. To assess the abilities of untrained humans for systems thinking, Dörner (Dörner 1990) has studied them when performing control tasks for dynamical systems comprising only a relatively small number of state variables (e.g., regulating the temperature of a cooling chamber). He found that only a rather small proportion of subjects was able to solve such tasks appropriately. For the large majority, typical dynamical systems properties such as delayed or counterintuitive responses or interaction of control parameters posed strong difficulties precluding a successful solution. Since then many similar experiments have been carried out (for a survey, see e.g., the book of Sternberg and Frensch (Sternberg and Frensch 1991)), with similar findings.

Fortunately, our tendency to have only limited native talent for systems thinking can – at least to some extent – be compensated by mathematical systems modeling and analysis methods.

After having chosen a suitable state representation (which never is unique, since the same state information can always be expressed in different "coordinates"), the next core ingredient of a dynamical system is to model how the present state is affected by interactions within the system or between the system and its environment. This part of the model usually is referred to as the "state dynamics" of the system. For the massive body, state dynamics is given by the famous Newton's law (saying that change of momentum equals the sum of all external forces that act on the body, and that change of position is the momentum divided by the mass). In the general case, state dynamics is a specification of the change of state as time progresses.

This leads to two major types of models: the first type models time as progressing in discrete time steps. The dynamic law then specifies how the state $s(t)$ at the present time t can be obtained from the previous state $s(t-1)$ and all present influences ("inputs") $x(t)$ on the system:

$$s(t) = f(x(t), s(t-1)). \tag{5}$$

The second type of model views time as progressing continuously. In this case, one describes the state evolution by specifying the velocity ds/dt with which it changes:

$$\mathrm{d}s / \mathrm{d}t(t) = f(x(t), s(t)) \tag{6}$$

(note that the functions in both equations are generally not the same although we have given them the same name).

The difference between both formulations is that the second, continuous, formulation is closer to reality (since time flows continuously) and also often more apt to mathematical analysis; however, there is the slight inconvenience that it requires to "integrate" the state velocity in order to arrive at the state itself. The discrete formulation is more convenient in this regard; it is also more apt for computer simulation, which are inherently discrete, and it can be used to construct arbitrarily close approximations to a given continuous formulation by choosing sufficiently small units for time steps.

Obviously, both types of models use the previously considered "mapping models" as the "vehicle" to represent state evolution. Therefore, many of the modeling techniques developed for mapping models are also useful for constructing dynamical systems and the major linear mapping model types give rise to corresponding taxonomies of dynamical systems.

Often, the state $s(t)$ is directly identified with the output $y(t)$ of the dynamical system model. However, in the general case the focus of interest may be some function $y(t) = g(x(t), s(t))$ of the state and the input variables. This brings in another mapping model. However, this does not affect the state evolution, it just acts as a kind of "feature extraction" (namely, forming the "feature" y) from the current situation (x, s).

6 Some Examples of Dynamical Systems

Their versatility makes dynamical systems models widespread in many areas. A particularly important class of dynamical systems is characterized by linearity in their state evolution function, i.e.,

$$\mathrm{d}s / \mathrm{d}t(t) = \mathrm{A}s(t) + \mathrm{B}x(t). \tag{7}$$

Here, we have used bold letters to indicate that usually A, B are matrices and $s(t)$ and $x(t)$ vectors that contain more than a single variable.

Since many phenomena behave linearly "in the small," models of the type (7) are applicable in many situations. They can describe such diverse phenomena as the behavior of particles, learning in neural networks in linearized approximation, vibrations in materials, the time course of temperature in heated bodies, or the dynamics in populations when they are far from saturation.

An important property of linear dynamical systems – inherited from linear mappings – is the constructability of their solutions as a weighted superposition of a finite set of "basis solutions." This makes them much easier to deal with than more general, nonlinear systems. However, linearity of a dynamical system *does by no means imply linearity of the state evolution*. For instance, the simplest linear differential equation

$$ds / dt(t) = \lambda s(t), \tag{8}$$

(which is perhaps the "most famous differential equation" at all) states that the rate of change of $s(t)$ is linearly proportional to its magnitude. This is a natural characteristic for many processes, since in the absence of external factors, growth always tends to be proportional to what is growing, no matter whether it is money in a bank account, the size of a bacteria culture, the number of light photons in a laser medium, or the number of links pointing to an internet hub. However, the well-known solution of the linear dynamical system (8) is a *regular typography* behavior, namely exponential growth (or decay, if $a < 0$) $s(t) \propto \exp(\lambda t)$.

Exponential growth sooner or later hits limits caused by the finiteness of all that matters. Modeling such "saturation effects" destroys the linearity of the dynamical system (thereby enabling solutions that stay within finite limits). For instance, the famous logistic equation

$$s(t+1) = a \cdot s\,(t) \cdot (1 - s(t)) \tag{9}$$

is approximately linear as long as $s(t) << 1$, leading to exponential behavior (each time step the magnitude of $s(t)$ grows approximately by a factor of a.[1]

Provided $a > 1$, sooner or later $s(t)$ will no longer be small and – by virtue of the factor $(1 - s(t))$ – start to reduce (or even reverse) the effect of the "growth factor" a.

Feigenbaum (Feigenbaum 1980) unraveled the surprisingly rich dynamic behavior that arises as a consequence of this innocent-looking nonlinearity: for some values of a, the system state converges to certain stationary values. For other values, the state begins to approach a "limit cycle," i.e., more and more closely periodically cycles between a finite number of values, where the period (the number of time steps until a cycle repeats) depends very sensitively on the value of a and can become arbitrarily large. Finally, for some values of the behavior of $s(t)$ exhibits what is termed "deterministic chaos," producing a highly irregular though deterministic sequence. While the nonlinearity ensures that $s(t)$ always remains bounded, small perturbations to its value at some time t get exponentially magnified in the number of time steps that elapse.

While many of these phenomena have been known to mathematicians already for quite some time, the work of Feigenbaum sparked a strong interest in these phenomena also outside of mathematics and contributed strongly to the spreading of new thought patterns: self-similarity as a parsimonious principle to specify structure across many scales, deterministic chaos as a way to grasp how highly sensitive behavior can be reconciled with rigid rules, and cascaded bifurcations as a "road" connecting orderly behavior and chaos.

Neural networks (Bishop 2006; Ripley 1996) dynamical systems: here the neural activity response $s_r(t)$ of a neuron at some location r can be modeled as a nonlinear function of the weighted sum of the activities of all other neurons at the previous time step

[1] Note that this time we are using discrete time steps and, therefore, have the "difference equation" analog of the continuous-time differential equation (8).

$$s_r(t) = \sigma\left(\sum_{r'} w_{rr'} s_{r'}(t-1)\right). \tag{10}$$

This is a nonlinear dynamical system with as many coupled equations as there are neurons. The nonlinear function $\sigma()$ describes how much a neuron is activated through the summed activities of all neurons connected to it. The "synaptic weights" $w_{rr'}$ specify the sign (excitatory or inhibitory) and strength of these influences and largely determine the dynamic characteristics of the dynamical system. In view of the great richness of already the single logistic model above it may not be too surprising that the model equations (11) can exhibit an even much richer behavior (Haschke and Steil 2005) which can by suitable choices of the $w_{rr'}$ couplings not only produce all kinds of chaotic behavior but – at least in principle – any computation that is computable.

In a way, the state $s_r(t)$ in the above model can be viewed as not only a time-dependent, but also as a space-dependent variable. We can make this more explicit by the notation $s_r(t) = s(r, t)$. Viewing r then as a continuous variable leads to dynamical systems models for continuous media, such temperature flow in a work piece, excitation in a layer of laterally interacting neurons, or the concentration of chemical substances in a reaction vessel. The space-variant state variables $s(r, t)$ are usually denoted as "fields," and their state equations then in addition to the time derivative also contain dependencies on their spatial derivatives ("gradients"). For instance, the rate of change of temperature at some point r is linearly proportional to the average difference to the temperatures in the immediate surround of that point. It can be shown that this "average local difference" is given by the second spatial derivatives, leading to the heat equation

$$\frac{\partial s(r,t)}{\partial t} = D\left(\frac{\partial^2 s(r,t)}{\partial x^2} + \frac{\partial^2 s(r,t)}{\partial y^2} + \frac{\partial^2 s(r,t)}{\partial z^2}\right) \tag{11}$$

describing the time evolution of the "temperature field" $s(r, t)$ of some body. Equation (11) is an example of a "prototypical" field model that applies in similar forms for a wide range of situations characterized by linear transport phenomena.

The introduction of dependencies on the spatial derivatives of the "fields" leads from differential to *partial differential equations* (again with time- and space-discrete counterparts for computer simulation).

Again, a major distinction is between linear and nonlinear equations (linearity being defined with respect to all dependencies that involve the state or one of its derivatives). For the linear type, an important fact is that that solutions can again be generated as superposition from basis solutions, but generally everything gets a bit more complicated, in particular when nonlinearity is involved.

A famous class of two-dimensional nonlinear dynamic (field-) systems comprises the so-called Turing systems (Turing 1990). They describe the concentrations of two chemically interacting substances which also undergo spatial diffusion with different diffusion coefficients. Already their inventor, Alan Turing, showed that these systems can exhibit sophisticated spatio–temporal pattern formation, which made them an important tool to model various structure formation processes observed in

chemistry or biology. With the hindsight of today it is highly remarkable that Turing not only created the theoretical basis of digital computing, but also the basis of computation and structure formation through continuous, spatio–temporal dynamical systems.

Innocent-looking examples of great dynamic richness are the Selkov equations:

$$\mathrm{d}U/\mathrm{d}t = \mathrm{D}_u \Delta U - UV^2 + F(1 - U),$$

$$\mathrm{d}V/\mathrm{d}t = \mathrm{D}_v \Delta V - UV^2 + (F + k)V.$$

For different values of control parameters (r, s), these equations exhibit very different spatio-temporal pattern morphologies, ranging from various dot- and stripe patterns to patterns resembling biological cell division processes and more (Pearson 1993).

Another kind of equation results if we replace the spatial derivatives by integrals. Such spatio-temporal integral equations can summarize the dynamics of neural layers due to lateral neuron interactions. These "neural field models" have been used, e.g., to model the dynamics of localized responses of neural layers to input stimuli. By making the lateral interactions adaptive and their changes activity dependent, these neural fields offer a model for the formation of topographic brain maps (Suder, Worgötter and Wennekers 2001). A highly abstracted and computationally simplified version of these models has become known as the "self-organizing map" (Kohonen 2001; Ritter et al. 1992) and has become useful not only in brain modeling but also as a tool in many applications.

7 From Deterministic to Stochastic Models

The previous models were fully *deterministic*: given their inputs, they yield an unambiguous and deterministic answer.

Many real-world phenomena behave differently: even a thrown dice – although its motion is governed by the deterministic laws of classical physics – confronts us with a result that is nondeterministic in practice and that can only be described by statistical means.

The stochasticity of most events in our environment must have had a major impact on brain evolution. Discovering patterns in noise is without doubt a major challenge to brain circuits. Computer vision and speech recognition confront us with similar challenges at a technical level. Stochastic models are likely to be the "common currency" that we have to develop in order to make progress with both modeling brain function and advancing the state of the art in technical pattern recognition and analysis systems.

However, these efforts have revealed that noise is not always only a nuisance: in some circumstances, noise can also be beneficial: optimization algorithms in high-dimensional spaces can benefit from the injection of noise to reduce the risk of becoming trapped in unfavorable local minima. Zero-sum games admitting no consistent strategy (absence of a "Nash-equilibrium") allow consistent "mixed strategies" in which "pure" strategies are selected according to a suitable random distribution.

In many cases (including the dice example) the reason for apparent nondeterminism is the well-known ability of microscopic and essentially unpredictable effects to have under many real-world conditions a macroscopic effect on the observed outcomes.

Since in most cases microscopic effects cannot be modeled themselves (at least, within reasonable computational costs), it is necessary at least to model their macroscopic effects on the observable outcomes. This leads to stochastic modeling frameworks (Kalin and Taylor 1998), in which "ordinary" variables are complemented with "random variables." Loosely speaking, such "random variables" can be thought of as variables whose value can fluctuate randomly, but in accordance with a specified probability density that characterizes the random variable. The resulting models are also termed generative models, since they describe data as "generated" from a process in which random variables are involved.

The simplest approach is to model the stochastic effects as an additive superposition of a random variable on the result of a deterministic mapping function, leading to

$$y = f(x,\theta) + \eta. \tag{12}$$

Here, η denotes the new random variable. A frequent assumption then is to assume that η follows a Gaussian distribution with zero mean and fixed standard deviation σ.

This leads to models for the optimization of which still many well-established methods are available.

Analogous extensions apply to dynamical systems (but we refrain from writing down the resulting equations, which look similar to (5) and (6).

8 Coping with Uncertainty

Historically earlier, in the attempt of modeling and analyzing games of luck, mathematicians have developed the concept of probability. It first arose as a kind of "limit" of the relative frequency of an event (for instance, the relative frequency of head coin tosses when the number of throws goes to infinity). this "frequentist view" later has become complemented by what is now called the "Bayesian view" which views a probability as a direct expression of our belief about the happening of an event (Bartholomew 1965). For instance, if the probability of an event is zero or one, we have complete certainty about the absence or presence of that event,[2] while complete uncertainty is represented by equal probabilities (e.g., 1/6 for a particular dice value) for all possible outcomes of an event.

Probabilities allow to "make uncertainty precise": the Shannon entropy

$$S = \sum_i p_i \log(1/p_i) \tag{13}$$

[2] We omit the discussion of measure-theoretic fine points admitting the happening or nonhappening of events even if their probability is zero or one, as long as such "exceptions" are sufficiently seldom so as not to affect any "expectation values," such as averages etc.

provides a quantitative measure of the "amount" of uncertainty associated with a set of probabilities $p_1, p_2, \ldots p_n$ for a number of distinguishable outcomes $i = 1, 2, \ldots n$ of an "experiment" (Khinchin 1957). It can be directly related to a minimal number of unbiased yes–no questions whose missing answers are equivalent to the uncertainty that is associated with the probabilistic outcome.

From the perspective of thinking, probabilities provide a remarkable extension of our thinking patterns: it has been observed that humans are much better in drawing inferences from situations that are characterized through conditions composed of conjunctions ("round" and "fruit" and "red"), while we find it hard to imagine or draw inferences from situations characterized by disjunctions ("round" or "fruit" or "red") (Gutierrez et al. 2001). Describing how outcomes are likely to be distributed among alternatives, the semantics of probabilities fits the disjunctive pattern. This may explain why we find it hard to apply them correctly on an intuitive basis. On the other hand, targetting a "weak spot" of human reasoning, probabilities can be extremely effective in compensating this weakness when we apply them in the framework of probability calculus.

For instance, quantifying uncertainty by entropy can lend a principled basis to thinking heuristics that guide our direction of search: often, the largest gains are obtained when probing where our uncertainty is large. Using probability models, this heuristics can be made more concise in several ways. If probing depends on choosing an action and we have a probability model predicting the possible outcomes after the possible actions we can determine that particular action that leads to the smallest a-posteriori uncertainty. "Active learning" adopts this idea to request training examples maximizing expected learning progress (Hasenjäger and Ritter 1998). The idea of "committee machines" identifies optimal questions as those for which a "committee" of predictors is maximally split about the outcome (Freund et al. 1997). And the method of optimal experimental design (Chaloner and Verdinelli 1995) chooses maximal expected entropy reduction as a design guideline for the planning and parametrization of experiments.

9 Optimal Inference

An associated question is how to combine different "states of knowledge" in a proper way that correctly "propagates" the associated uncertainties. The fundamental "law" has first been formulated by the British mathematician Thomas Bayes (1702–1761) and is now known as "Bayes rule." Basically, it tells us how the knowledge of the occurence of some "context" c changes the probabilities $p(x)$ associated to the outcomes x of some other event into new, modified values $p(x|c)$ that reflect our new state of knowledge:

$$p(x|c) = N\, p(c|x) p(x), \tag{14}$$

i.e., the necessary "correction factor" $p(c|x)$ is – apart from a normalization factor N that has to ensure that the new probabilities again add up to 1 – just the probability

of the reverse situation: namely the probability of occurence of the context c given that outcome x is known to have happened.

This innocent-looking rule is the basis of what is called "optimal Bayesian inference" – drawing "optimal" conclusions in the presence of uncertainty, i.e., knowledge only about probabilities for events and contexts (Mackay 2003; Berger 1985).

For instance, in a medical application x might represent a number of possible illnesses, the "context" c would be a set of diagnostic symptoms, $p(x)$ would be the "a-priori" probability of illness x in the absence of any further information, and $p(x|c)$ would be the revised probabilities, knowing that the diagnostic symptom c is observed. In this case, Bayes rule is seen to "convert" knowledge about the probabilities $p(c|x)$ of symptoms in the presence of an illness x into knowledge $p(x|c)$ about the occurence of illness x when observing symptoms c.

Knowing $p(x|c)$ then provides us with the basis for making an optimal prediction about the outcome x (e.g., the illness) from the observed context or features c (e.g., the symptoms): this is the task of *classification*. If one associates misclassifications with a certain "cost," then, in the simplest case when all misclassifications carry the same cost, the optimal classification scheme should minimize the to-be-expected costs associated with the uncertainty. Such reasoning then leads to the simple and intuitive result that the optimal classification is to vote for the particular outcome x for which $p(x|c)$ is maximal (as compared to any $p(x'|c)$ of a "competing" outcome x'). Using this criterion is known as "optimal Bayesian classification," and (under the criterion of cost minimization) there is no better decision scheme possible.

The predictions made by Bayesian modeling can appear rather counter-intuitive to human thinking: assume observing a symptom c for a disease x, knowing that the symptom is associated with 99% of all cases of the disease (i.e., $p(c|x) = 0.99$), while the remaining 1% occurrences follow no particular pattern. To most people the observation of such a symptom would appear as strong evidence for the presence of the disease. Yet if the disease is very rare, e.g., having an a-priori probability $p(x) = 0.0001$, the average number of truly diagnosed positives in one million people will be 99 cases (99% of 0.0001×1 Million), while the remaining 1% occurences of the symptom that are unrelated to the disease will produce an average of 9,999 false alarms (1% of 999901). Under these conditions, less than 1% of the alarms (99 of 10,098) will actually indicate the presence of the disease, while most alarms will be totally unrelated with the disease.

Examples like these show that human thinking is sometimes far from optimal inference and can benefit significantly from using analytical models.

The conversion of symptoms into a "better"[3] probability distribution can be repeated: Bayes rule is then used as a principled way to include the information from additional observations, e.g., about further symptoms, into the shape of the probability distribution for the different illnesses. This allows at each point of the process to

[3] It may also happen that the new probability distribution $p(x|c)$ is more spread out than the "old" one $p(x)$: this possibility reflects the fact that a symptom can also be "confusing."

know which illness has the highest probability, given the current state of knowledge. In addition, the summed probabilities of all the other illnesses directly provide the probability of error, namely that the maximum probability illness happens to be the wrong diagnosis.

This framework of Bayesian inference has become developed into a powerful modeling framework. An important refinement has been the use of methods from graph theory to decompose the often high dimensional probability functions into products of lower dimensional factors. These *graphical models*(Lauritzen 1996) now offer a useful toolset for modeling statistical relationships and performing inference in a close-to-optimal (since usually approximations are unavoidable to perform the calculations) way.

10 From Modeling to Bayesian Learning

Using Bayesian inference to combine knowledge for reducing the uncertainty about events or outcomes can also be applied to the process of modeling itself: the uncertainty about the "correct" model becomes expressible as a probability distribution in the space of models, i.e., each model that is a priori a feasable solution candidate becomes assigned a probability value which measures the likelihood that it is the true model for the data at hand. The entropy of this probability distribution is then a direct measure of the total modeling uncertainty, and Bayesian inference can be used to recompute the probability distribution over models when new data become available and to keep track of the most probable model (Mackay 2003), much in the way as discussed previously in the illness example. Note also that this way of representing model uncertainty is different from modeling uncertainty within a phenomenon: modeling a coin toss uses probabilities to model the uncertainty in the tossing outcome. Our model of the coin toss may be formulated at a different level (e.g., the mechanics of the tossing etc.) and there will be some uncertainty about the model parameters, which is at an entirely different level than the tossing outcome.

The data-driven reshaping of our uncertainty about the "true" model for some set of data relationships can be viewed as a kind of learning. By its underlying framework, this type of learning has become termed *Bayesian learning*, and the mechanism of Bayesian inference has given rise to *Bayesian learning theory*, which is now one of the best developed theories of learning with applications in numerous fields.

However, for practical applications the space of models for almost any situation of interest is almost prohibitively large: it is, for instance, enormously larger than the space of potential illnesses in a diagnostic situation.

Therefore, to apply Bayesian learning theory to practical problems, the set of models usually is represented by a "single" model, but with a set of parameters which allow to "tune" it to different data sets. This transforms the task from having to deal with a general set of models to the task of having to deal with a general set of parameters, a still difficult but much more feasible problem.

This leads back to the familiar tool of mapping functions $y = f(x; \theta)$ that contain some parameter(s) θ. But the somewhat naive notion of "fitting" such a model to given data now has become replaced by a deeper view: instead of just looking for a "best fit," Bayesian learning theory allows to employ probabilities to (1) express our uncertainty about the "best" model in a much more principled and gradual way than before (where the only alternatives where "we don't know" and "we have a best fit"), and (2) it offers optimal Bayesian inference to integrate new information.

Yet, even with parametrized models Bayesian learning is computationally a rather demanding approach. Therefore, in practice there exist a number of simplifications of Bayesian learning which are computationally much lighter and have become standard tools for modeling.

One major simplification is to focus only on the parameter value with maximal probability. However, in a high-dimensional space this parameter value need not be very "typical." In such cases, it may be a better choice to use the "average value." Again, this may be a poor choice (e.g., for an annullus-shaped distribution of parameter values).

11 Maximum Likelihood Principle

The high computational burdens (but also its strengths) of the Bayesian modeling approach can be traced to the ambitious goal of identifying "the full shape of uncertainty." Technically, this means that the Bayesian approach tries to identify a *random variable*, which is a much more sophisticated "object" (being associated with an entire probability distribution) than an ordinary parameter.

The resulting burden can be avoided (with some of the power sacrificed) by detaching the uncertainty representation from the to-be-identified object. This is the idea that is shared by several nonbayesian stochastic modeling approaches: these approaches represent model uncertainty entirely as a result of uncertainty in the to-be-modeled data (e.g., due to noise); beyond that, there is no room for any inherent "information uncertainty" in the modeler. As a result, the to-be-identified object now is a plain parameter value θ. The data uncertainty is expressed as a stochastic data model $y = f(x, \theta) + \eta$ as encountered before (Kay 1993).

Different criteria for "the proper" choice of θ, given the data, can then be postulated. One of the most widely used approaches is the "Maximum Likelihood Principle" ("ML"-principle, Bishop 2006; Edwards 1972). It argues that the observation of data z should lead us to adopt those parameter values θ that grant a maximal probability of our observation in our model.

Formally this is easily expressed as

$$\theta_{ML} = \arg\max_{\theta} p\,(z,\theta) \tag{15}$$

where $p(z, \theta)$ denotes the probability of seeing data $z = (x, y)$ for parameters θ in our model.

For an additive stochastic mapping model (12) the "likelihood function" $p(z, \theta)$ is rather simply related with the probability density $P(\eta)$ of the additive noise:

$$p(z;\theta) = P(y - f(x,\theta)).$$

Plugging this into (15) directly leads to a (usually highly nonlinear) optimization problem which can have numerous local optima. Usually, one considers the equivalent maximization of the (somewhat "smoother") logarithm of the likelihood and special algorithms (most notably the so-called EM-method) have been developed for solving this task.

A major assumption is statistical independence of all data points. In this case, $p(z; \theta)$ becomes a product $\Pi p(z_i, \theta)$ of lower-dimensional single data point probabilities $p(z_i, \theta)$, and the logarithm of the likelihood function turns into the sum

$$L = \sum_i \log P(y_i - f(x_i,\theta))$$

12 Learning, Optimization and Risk Minimization

In this context, the frequently used least square error function

$$E = \sum_i (y_i - f(x_i,\theta))^2$$

minimized for determining optimal fitting parameters θ for a parametrized mapping model $f(x; \theta)$ can be recognized as a special case of the ML-principle, namely when the additive noise distribution in the stochastic data model is a Gaussian: in that case, $\log P(y_i - f(x_i, \theta)) \propto -(y_i - f(x_i, \theta))^2$ and *maximization* of L leads to the same solutions θ as *minimization* of the square error function E.

On should note that both the ML-framework and the Bayesian framework are conceptually different: the ML framework focuses on modeling data uncertainty and identifies a model that makes the observed data maximally likely. The Bayesian framework focuses on modeling *the entire modeling uncertainty* and finds a probability distribution across models that represents our state of modeling uncertainty. This makes it more comprehensive, since by construction it includes any "data parameters" and identifies corresponding probability distributions reflecting the information in the data about them. But the price to pay is a significantly higher computational burden, often necessitating severe approximations that preclude a simple answer to what is the best method in any case.

From a more computational perspective, both lines of approach lead to a similar task, namely minimizing a (usually) high-dimensional error or cost function. This has been generalized, viewing modeling as the minimization of a certain "risk" when acting according to an adopted model.

This view has become most prevalent in current research, since it offers a common framework for probabilistic (such as Bayesian and ML) and nonprobabilistic frameworks. Modeling in this case is formulated as the task of finding that particular model which minimizes a prespecified (domain-dependent) "Risk function" (Vapnik 1995).

Usually, the competing models are thought as selected through a parameter w and the risk function is assumed to be an additive sum of contributions $e(z_i; \theta)$ arising from the individual data points $z_i = (x_i, y_i)$ (for instance, in the case of (12) $e(z, \theta) = -\log(y_i - f(x_i; \theta))^2$).

Usually, this optimization cannot be done in a single step but must be performed iteratively. Each iteration step can be interpreted as a partial reduction of the total "model risk" (in the Bayesian case, "risk" would be a measure of model uncertainty, in the ML case of model mismatch). Downhill gradient following

$$\Delta\theta = -\varepsilon \cdot \partial E/\partial w(z,\theta) \tag{16}$$

is a rather general and intuitive procedure to change the model parameters in a step-wise fashion towards the (nearest local) risk minimum. In the case of a datapoint-wise additive risk function, the above equation can be approximated by cycling through all data points, changing the parameters at each step only by the part

$$\Delta\theta = -\varepsilon \cdot \partial e/\partial w(z_i,\theta) \tag{17}$$

that arises from the data point's risk contribution $e(z_i, \theta)$.

This allows to view each data point as giving rise to a "learning step" given by (17) (it turns out, that the cycling order through the data points is not very critical and can even be chosen at random under mild technical conditions).

13 Bias, Complexity, and Generalization

While we often may view modeling as a construction process, a logically equivalent view is to consider the set F of all models that we a-priori deem as conceivable for a phenomenon of interest, and to use data to narrow this (usually very large) set to the smaller subset of models that match the data well. In this view, machine learning appears as the process to achieve this narrowing, e.g., by retaining only those models θ from F whose risk function $E(\theta)$ has a small value.[4]

This has an informal analogy in thinking: successful thinking also depends to a considerable extent on the finding of "insightful perspectives" on the situation at hand. This is a combination of framework selection, making the right abstractions and recognizing a manageable set of factors whose combination carry the essential part of the problem. In its entirety, this is a highly complex process that involves experience and mental skill in ways that are only little understood to date. The well-formalized framework of model selection in machine learning may offer a highly simplified scenario in which some of the major constraining factors become visible and analysable: the role of model complexity and data, the various methods to narrow the range of viable models, and the trade-offs with regard to generalization.

[4] We intentionally use for the models the same letter θ that we previously used for the model parameters to indicate that we mostly think of parameterized models, for which model identity and the associated parameter value can be identified.

From the perspective of this much more constrained framework the key question of modeling is: How well do the retained models *describe new data*, i.e., data that were not used for the narrowing step? To measure this *generalization ability* therefore requires to split the available data into a "training data set" (used for the narrowing step) and a "test data set" used to evaluate the retained models.

It is a major difficulty that in general the set of retained models (characterized by low risk values on the training set) can differ strongly on the test data set, i.e., exhibit widely varying generalization ability. This happens particularly when the initial model set F has been "too large." For instance, when F is the set of "smooth" functions, then F contains very many functions that fit any given number of training data points perfectly, but still can assume arbitrary values on new data points. As a result, even a large training set cannot single out a retained set in which all members exhibit good generalization. The retained set always suffers from a large variance of its members, degrading generalization to small values.

On the other hand, choosing F as the set of linear functions, already two data points are sufficient to narrow F to a single remaining candidate (if we require zero risk). If the underlying phenomenon is indeed linear (or close to linear), this candidate will exhibit a good generalization on new data points. Therefore, the strong bias of a "small" F has – on the one hand – a distorting effect, trying to impose some structure on the data. On the other hand, this distorting effect also facilitates good generalization – if the bias is in harmony with the phenomenon under consideration. Reducing the bias reduces the distortion, but also the generalization by reducing the variance within the retained set. Finding the best balance is known as the problem of the "Bias-Variance-Dilemma."

Statistical machine learning theory (Vapnik 1995) has analyzed this situation very closely, with the major insight that the decisive factor for good generalization is the relation between the "richness" of the chosen model family F and the richness of the available training data. If F is very rich, i.e., contains also very "complex" models, generalization performance tends to be low, unless the higher richness of models is balanced by a correspondingly larger data set. And the example of smooth functions showed that it is easy to specify model sets for which no finite training data set can ensure good generalization.

To be able to choose the right model complexity on a principled basis, different measures of richness have been developed, ranging from the entropy of F as the most detailed measure, to various "capacity dimensions," such as the Vapnik-Chervonenkis (VC-) dimension, that are more practical to compute for actual problems. All these measures allow attaching complexity measures to model sets from which one can get (probabilistic) upper bounds on the generalization error. However, these bounds are usually rather pessimistic.

For practical purposes, the method of choice remains to use the training data to construct a model and to evaluate its generalization performance on an independent test data set, being aware that this provides just an estimate, which, however, often is accurate enough. Model selection then usually is based on a 3-splitting of the data into training, test and validation data sets. The first two are used to estimate generalization for a particular candidate F. The third is used to decide among competing

F of different "richness." Usually, the competing function sets form a nesting sequence, with richer function sets encompassing less rich ones. A frequent axis of variation is the number of parameters of the models in F: the more parameters the richer is F. In the case of neural networks, the parameters are the connection weights and the parameter axis becomes essentially the size of the network.

14 Limits of Modeling

Models always have to be constructed from a combination of prior knowledge and data. More often than not, this information is scarce or uncertain, limiting the ambitions of the model builder. For a number of modeling scenarios, these limitations can be analyzed sufficiently thoroughly to allow the specification of mathematical bounds of what is achievable and what is not (or, if it turns out to be achieved, is due to luck).

When random noise has a strong influence, there is always a competition between "patterns" that occur purely due to chance, and systematic patterns that have their origin in a real phenomenon. Distinguishing both cases is a common task for anybody attempting to model relationships in observed data. It has become good scientific practice to perform this assessment in terms of *significance values*. The underlying idea is to state the absence of the phenomenon of interest as a "null hypothesis" and then use this as the basis to compute the probability p that the observed candidate for a systematic pattern can have been generated purely by chance. If the resulting "p-value" is below a specified threshold (e.g., 0.01), then one has good reason to assume that the null hypothesis is unlikely to have produced the observed "pattern candidate" and the pattern is likely to reflect a real phenomenon.

Another question is to ask for bounds on the reliability of predictions that a model can make. Classification is a major setting where this question can be analyzed rather completely, yielding rigorous performance bounds in terms of the involved class probabilities.

Given that we have this "maximal" state of information (i.e., knowing all the class probabilities $p(x|c)$) of information, any "overlap" between the densities of different classes precludes a perfect distinction of the underlying classes x from the features c alone, leaving an unavoidable classification error $e(x)$ which can be computed for each class x as

$$\mathrm{e}(x) = \sum_{x' \neq x} \int_{p(x'|c) > p(x|c)} p(x' \mid c)\, \mathrm{d}c.$$

This is the probability of the optimal classifier misclassifying an instance of class x as some other class x' due to $p(x|c)$ being smaller than some $p(x'|c)$ *although* x is from class c.

This is known as the "Bayesian classification limit" (Bishop 2006; Berger 1985) no model can do better than this. If this error is too high, the only remedy is to try to find a more discriminative feature set $\{c'\}$, giving rise to a new set of probability densities $p'(x|c')$ with reduced overlap and, thereby, a smaller Bayesian classification limit.

However, usually the Bayesian limit is not achievable in practice: usually the underlying probabilities $p(x|c)$ are not known exactly, but only as approximate estimates from available data.

This leads to the important question how to transform data and prior knowledge in an "optimal way" into the "most adequate" probability function (Mackay 2003). If we already have some probability function as a starting point, we have seen that Bayes Rule is the correct principle to integrate additional information from data. However, the question remains how to construct the initial "a-priori" distribution needed to get Bayes Rule started. the so-called "maximum entropy principle" (Jaynes 1957) provides a concise answer to that question: it basically states to select from the range of a-priori possible a-priori probability functions that particular function (1) which satisfies the given constraints (e.g., having a specified mean value and variance) and at the same time (2) has the maximal entropy of all its "competitors" satisfying (1).

This has an intuitive interpretation: all our certainty is in the given constraints and only in these. So we should not pick a probability distribution that represents less uncertainty, and, therefore, entropy, that is expressed in the constraints that we do know.

Frequently, identifying probability distributions is an overambitious modeling approach (Ripley 1996). If one is interested, for instance, in just identifying a best fitting model from a parametrized family, the ML principle offers a computationally manageable framework, determining the to-be-identified parameter values as the maximum of the likelihood function. This brings another fundamental modeling limit into view: namely, how accurately can the "true" parameter values be estimated from the available data? The *Cramer-Rao bound* (Kay 1993) answers this question, essentially stating that the expected variance in the estimation is bounded by the inverse of the local curvature of log-likelihood function at its maximum point. Intuitively, this means that the "sharpness" (measured in terms of local curvature) of the peak of the log-likelihood function determines how "sharply" the available data allows to estimate a (set of) parameters of interest.

These are only some of the major instruments to assess limits of modeling, and they all address limitations arising from the finiteness of available data. In practice, these limitations can become aggravated by limitations on the available computational resources: usually, the high dimensionality of the feature spaces involved in realistic situations precludes certain theoretically optimal computations (such as, e.g., high-dimensional integrations required for optimal Bayesian inference). As a result, these computations have to become replaced by more or less suboptimal approximations, leading to a correspondingly reduced prediction performance of the model. Although this "curse of dimensionality" becomes less acute with the steadily increasing performance of computer hardware, it will not go away, since many computations are exponentially expensive in the size of the problem. therefore, even when our computational resources are likely continue their exponential growth for some more years into the future, some ambitious goals such as noninvasive brain–computer interfaces that hinge on the detection of highly delicate patterns in strong noise, may forever be limited in the detail they can deliver – despite the demonstrated and useful ability to extract nontrivial coarse grained information (Nicolelis 2001) about brain contents.

15 Concluding Remarks

There is much more to say about modeling than would fit into this limited chapter. We have only touched upon a core set of modeling approaches, leaving out important methods such as rule based models to represent models as a collection of logical statements, fuzzy sets for coping with uncertainty in a computationally cheaper way than combining probability distributions, graphical models to represent the interaction of probabilistic causes, or game theory to model the effect of action decisions during the interaction between two or more agents.

Picking the adequate modeling framework therefore resembles often an art that can only be partially guided by first principles. It has to benefit substantially also from experience and not seldom some amount of trial-and-error, addressing mutually conflicting issues such as accuracy, availability of sufficient data, computational feasibility and adaptability to new situations.

Looking at the brain as our "best modeling engine," we also are led to the impression that our practiced approaches are likely to be still far from what can be extracted from data: although the brain is subject to the same fundamental limits exposed in the previous section as all our algorithms, it effortlessly extracts with a high reliability a huge range of rich structures from extremely high dimensional data consisting of spike trains elicited from our sensing apparatus by scattered electromagnetic fields and pressure oscillations in the ambient air. The resulting "models" that we find in our consciousness are crisp and immediate, and they are the "color" in which our reality becomes "painted" every second. They allow us to make immediate and usually amazingly correct predictions for all our movements and activities, and it remains a major challenge to bring artifacts, such as robots, at least into remote proximity to what brains enable their owners to do. A second challenge is to transpose such capabilities into domains for which evolution had neither need nor the substrate to develop similar modeling powers. One major example is the internet as our most dynamically growing "information biotope," but also the distant micro- and macrocosms into which the ever more advances sensors of scientific inquiry connect our natural senses. A useful step in this direction might be "brain-adequate" human-machine interfaces (Ritter et al. 2006) to create better synergies between our natural pattern discrimination abilities and the computational abilities of machines.

On the other hand, the research associated with modeling has helped to pinpoint a number of questions that are also at the core of thinking and of a better understanding of "the mechanics of thinking": first of all, the richness of model types developed over time gives us a better appreciation in how many ways information can be represented and how these representations are related to each other. Perhaps even more important, models act as "information compressors," often capturing the essence of large data sets in a much smaller set of essential degrees of freedom and thereby suggesting new abstractions and structural relationships that may otherwise remain hidden (for a recent study showing that already information compression alone can induce interesting categorizations within data, see Heidemann and Ritter 2008).

The limitations apparently built into the human cognitive system (severely restricted short term memory, low absolute accuracy for memorized quantity) may even exist to enforce strongly compressed representations and thereby ensure more flexibility in the long run. Thirdly, the success of machine learning techniques and their intimate relation with model building have at least raised the hopes that some aspects of thinking and abstraction might be automatable. However, the richness of dynamical systems encountered and partly analyzed in theoretical and experimental work also should make us aware that we are very likely to encounter still many major surprises and challenges along our way towards a deep understanding and replication of nontrivial thinking processes.

On our way, we may have forgotten the ropes: light both ends of the first rope, and one end of the second rope. When the first rope has fallen into ashes, also light the nonburning end of the remaining, second rope.

References

Duch W, Wieczorek T, Biesiada J, Blachnik M (2004) Comparison of feature ranking methods based on information entropy. Proc IEEE Int Joint Conf Neural Netw 2:1415–1419

Alligood KT, Sauer TD, Yorke JA (2000) Chaos. An introduction to dynamical systems. Springer, Berlin

Bartholomew DJ (1965) A comparison of some bayesian and frequentist inferences. Biometrika 52(1/2):19–35

Berger J (1985) Statistical decision theory and Bayesian analysis. Springer, New York

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 10:273–304

Dörner D (1990) The logic of failure. Philos Trans R Soc Lond B 327:463–473

Edwards AWF (1972) Likelihood. Cambridge University Press, Cambridge

Feigenbaum MJ (1980) Universal behavior in nonlinear systems. Los Alamos Sci 1:4–27

Freund Y, Seung S, Shamir E, Tishby N (1997) Selective sampling using the query by committee algorithm. Mach Learn 28:133–168

Gallant AR (1986) Nonlinear statistical models. Wiley, New York

Gutierrez F, Garcia-Madruga JA, Moreno S, Carriedo N, Johnson-Laird PN (2001) Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. Q J Exp Psychol 54A:613–632

Harrel FE (2001) Regression modeling strategies, Springer Series in Statistics. Springer, Berlin

Haschke R, Steil J (2005) Input space bifurcation manifolds of recurrent neural networks. Neurocomputing 64:25–38

Hasenjäger M, Ritter H (1998) Active Learning with local models. Neural Process Lett 7:107–117

Heidemann G, Ritter H (2008) Compression for visual pattern recognition. IEEE ISCCSP Conf Proc:1520–1523

Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York

Jaeggi SM, Buschkuehl M, Jonides J, Perrig WJ (2008) Improving fluid intelligence with training on working memory. PNAS 105(19):6829–6833

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620–630

Jolliffe IT (2002) Principal component analysis, Springer Series in Statistics, 2nd edn. Springer, Berlin

Kaplan D, Glass L (1995) Understanding nonlinear dynamics. Springer, Berlin

Karlin S, Taylor HM (1998) An Introduction to stochastic modeling, 3rd edn. Academic, New York
Kay SM (1993) Fundamentals of statistical signal processing. Prentice Hall, New Jersey
Khinchin (1957) Mathematical foundations of information theory. Dover Books, New York
Kohonen T (2001) Self-organizing maps. Springer, Berlin
Lauritzen SL (1996) Graphical models. Oxford Statistical Science Series. Oxford University Press, Oxford
MacKay DJC (2003) Information theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge
Meinicke P, Klanke S, Memisevic R, Ritter H (2005) Principal surfaces from unsupervized kernel regression. IEEE PAMI 27:1379–1391
Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81–97
Minsky M, Papert SA (1988) Perceptrons: expanded edition. MIT, Cambridge, MA
Nicolelis M (2001) Actions from thoughts. Nature 409:403–407
Pearson JE (1993) Complex patterns in a simple system. Science 261:189–190
Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
Ritter H, Martinetz T, Schulten K (1992) Neural computation and self-oranizing maps. Addison Wesley, Boston, MA
Ritter H, Kaper M, Lenhardt A, Ontrup J (2006) Making human-machine interfaces more brain-adeqate. In Brain-inspired It III (International Congress Series), pp 15–21
Roweis S, Saul LK (2000) Nonlinear dimension reduction by locally linear embedding. Science 290:2323–2326
Schölkopf B, Smola AJ (2002) Learning with kernels. MIT, Cambridge, MA
Shawe-Taylor J, Cristianini N (2000) Support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
Special Issue on Variable and Feature Selection. J Mach Learn Res 3 (2003)
Sternberg RJ, Frensch PA (1991) Complex problem solving: principles and mechanisms. Lawrence Erlbaum, New Jersey
Suder K, Wörgötter F, Wennekers T (2001) Neural field model of receptive field restructuring in primary visual cortex. Neural Comput 13:139–159
Turing AM (1990) The chemical basis of morphogenesis. Bull Math Biol 52(1–2):153–197 (reprint of the original 1953 paper)
Vapnik VN (1995) The nature of statistical learning theory. Springer, Berlin
Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Fransisco

Index

B

T