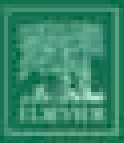
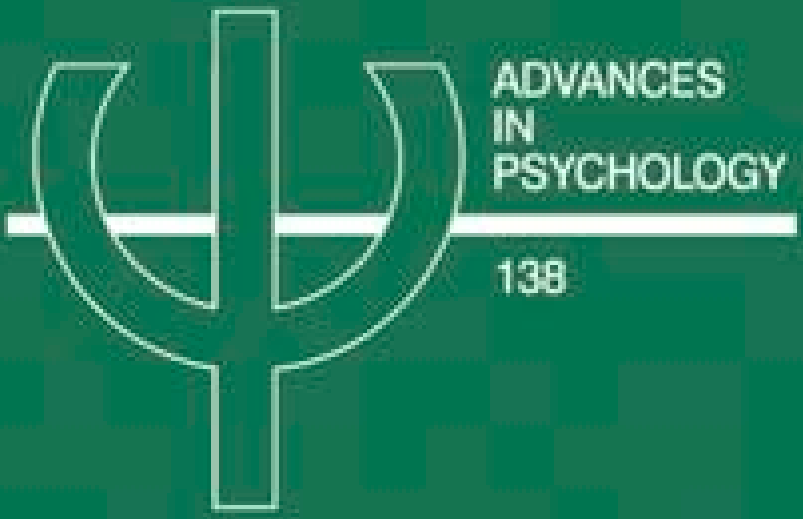




Mental Models and the Mind

Current Developments in Cognitive Psychology, Neuroscience, and Philosophy of Mind

Carsten Held
Markus Knauff
Gottfried Vosgerau
Editors

MENTAL MODELS AND THE MIND

Current Developments in Cognitive Psychology, Neuroscience, and Philosophy of Mind

ADVANCES IN PSYCHOLOGY

138

Editor:

G. E. STELMACH

ELSEVIER

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

MENTAL MODELS AND THE MIND

Current Developments in Cognitive Psychology, Neuroscience, and Philosophy of Mind

Edited by

Carsten HELD
Universität Erfurt, Germany

Markus KNAUFF
Max-Planck-Institute, Germany

Gottfried VOSGERAU
Eberhard-Karls-Universität, Germany

ELSEVIER

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER B.V.
Radarweg 29
P.O. Box 211
1000 AE Amsterdam
The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego
CA 92101-4495
USA

ELSEVIER Ltd.
The Boulevard
Langford Lane, Kidlington
Oxford OX5 1GB
UK

ELSEVIER Ltd.
84 Theobalds Road
London
WC1X 8RR
UK

First edition 2006

Library of Congress Cataloging in Publication Data
A catalog record is available from the Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN-13: 978-0-444-52079-1
ISBN-10: 0-444-52079-1
ISSN: 0166-4115 (Series)

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

Contents

Part III: Perception, Emotion, and Language

Part IV: Philosophy of Mind

Preface

The present book is the result of a workshop on "Mental Models and the Mind" that has been held at the University of Freiburg in the summer 2003. The workshop brought together researchers from a variety of disciplines: Cognitive psychologists reported their research on the representation and processing of mental models in human memory. Cognitive neuroscientists demonstrated how visual and spatial mental models are processed in the brain and which neural processes underlie visual and spatial thinking. Philosophers talked about the role of mental models in relation to perception, emotion, representation, and intentionality. Computer and education scientists reflected on the importance of mental models, both theoretically and application-driven. As it is often the case after a stimulating workshop, the idea of a book publication based on the contributions quickly arose. We have asked all workshop participants for extended versions of their papers and have invited other colleagues to contribute to the book. We owe special thanks to Phil Johnson-Laird for his support and for an original contribution.

We gratefully acknowledge financial support by several organizations for the workshop and for our own research, including the writing and editing of this book. The Freiburg workshop was supported by the *Fritz-Thyssen-Stiftung* within the *Cross Section Area: Image and Imagery.* Our work has also been supported by the *Deutsche Forschungsgemeinschaft* (DFG) through the Transregional Collaborative Research Center *SFB/TR 8 Spatial Cognition* and by the *VolkswagenStiftung* through the Research Project *Self-consciousness and concept formation in humans.* Markus Knauff is supported by a Heisenberg Award from the DFG. Gottfried Vosgerau has been supported by a DAAD exchange fellowship held at NYU and a dissertation fellowship of the *Studienstiftung des deutschen Volkes.*

Every article in this book has been reviewed by another contributor and by one external colleague. These colleagues provided authors and editors with helpful comments and suggestions, in particular with respect to the multi-disciplinary audience we hope to reach. We thus owe thanks to all our authors and the following external reviewers: Wolfgang Huemer, Georg Jahn, Christoph Klauer, Albert Newen, Klaus Oberauer, Wolfgang Schnotz, Walter Schaeken, Bernhard Schröder, Ralph Schumacher, Gerhard Strube, Kai Vogeley, Lara Webber, Stefan Wölfl, and Hubert Zimmer. Thanks also to Nadine Becker, Steffi von dem Fange, and Doreen Schmidt for technical support, and, on the publisher's side, to Fiona Barron, Joyce Happee, and Simon Pepping for a smooth and effective collaboration.

Carsten Held, Markus Knauff, Gottfried Vosgerau
September 2005

This Page is Intentionally Left Blank

Contributors

Vicky Dierckx:
Department of Experimental Psychology, Ghent University
Henri Dunantlaan 2, 9000 Ghent, Belgium

Vinod Goel:
Department of Psychology, York University
Toronto, Ontario, Canada M3J 1P3
vgoel@yorku.ca

Verena Gottschling:
Philosophisches Seminar, Johannes-Gutenberg-Universität
Jakob-Welder-Weg 18, 55099 Mainz, Germany
gottschl@uni-mainz.de

Carsten Held:
Wissenschaftsphilosophie, Universität Erfurt
99105 Erfurt, Germany
Carsten.held@uni-erfurt.de

Barbara Hemforth:
Laboratoire Parole et Langage, Université de Provence
29, avenue Robert Schuman, 13621 Aix-en-Provence Cedex 1, France
barbara.hemforth@lpl.univ-aix.fr

Philip N. Johnson-Laird:
Department of Psychology, Princeton University
Princeton, NJ 08544, USA
phil@princeton.edu

Lars Konieczny:
Center for Cognitive Science, Albert-Ludwigs-Universität
Friedrichstr. 50, 79098 Freiburg, Germany
lars@cognition.uni-freiburg.de

Markus Knauff:
Center for Cognitive Science, Albert-Ludwigs-Universität
Friedrichstr. 50, 79098 Freiburg, Germany
knauff@cognition.iig.uni-freiburg.de

Michael Pauen:
Institut für Philosophie, Otto-von-Guericke-Universität
Virchowstr. 24, 39016 Magdeburg, Germany
m@pauen.com

Klaus Rehkämper:
Institut für Philosophie, Carl-von-Ossietzky-Universität
26129 Oldenburg, Germany
klaus.rehkaemper@uni-oldenburg.de

Norbert M. Seel:
Institut für Erziehungswissenschaft, Albert-Ludwigs-Universität
79085 Freiburg, Germany
seel@uni-freiburg.de

Hannelore Van der Beken:
Department of Experimental Psychology, Ghent University
Henri Dunantlaan 2, 9000 Ghent, Belgium

André Vandierendonck:
Department of Experimental Psychology, Ghent University
Henri Dunantlaan 2, 9000 Ghent, Belgium
Andre.Vandierendonck@UGent.be

Gottfried Vosgerau:
Philosophisches Seminar, Eberhard-Karls-Universität
Bursagassse 1, 72070 Tübingen, Germany
vosgerau@uni-tuebingen.de

Randall L. Waechter:
Department of Psychology, York University
Toronto, Ontario, Canada M3J 1P3

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

General Introduction: Current Developments in Cognitive Psychology, Neuroscience, and the Philosophy of Mind

"Cognitive psychology," "cognitive neuroscience," and "philosophy of mind" are names for three very different scientific fields, but they label aspects of the same scientific ambition: to understand the nature of mental phenomena. Cognitive psychologists study mental processes as they are indispensable for understanding human experience and behavior. They systematically observe such behavior and then draw inferences from the observed data about unobservable mental processes. They also apply their results to various domains of human life, including the design of new teaching methods and the treatment of mental illness. Cognitive neuroscientists are concerned with the connection between mental processes and the brain. They investigate how brain-events affect human behavior. Philosophers of mind study the nature of mind, including consciousness, mental representation, and rationality. They ask questions such as: What is the relation between mind and brain, on the one hand, and mind and world, on the other? Can machines think? How is the realm of beliefs and knowledge connected to behavior? How come can I think about my own mental states? For many decades, the three disciplines worked in relative isolation, but today they strongly overlap under the roof of cognitive science. The goal of modern cognitive science, from our point of view, is to explain how cognitive processes are related to and can be measured via behavior, how they are computationally realized, and how these computations are biologically implemented in the brain.

In all sub-fields of cognitive science, the vast majority of researchers are familiar with the term "mental model." Sometimes the expression is used as a synonym for "mental representation," but in most areas it has a more

precise meaning. Building a bridge between the philosophy of mind and the empirical sciences of the mind/brain, the present book develops a new perspective on the concept of mental models—from the points of view of the mentioned disciplines: cognitive psychology, cognitive neuroscience, and philosophy of mind. In the following, we provide a short introduction to the field. We initially sketch some history of cognitive psychology and the newly emerging cognitive science as the background against which the conception of a mental model has been invented (sec. 1–3). Then we describe the conception in more detail (sec. 4) and outline the neuroscientific research it has inspired in recent years (sec. 5). The last three sections draw a connection to the philosophy of mind. Mental models are a special kind of mental representation. We sketch what philosophers think about mental representation (sec. 6) and about how to make it scientifically accessible via a procedure called "naturalization" (sec. 7). This will prepare the ground for a short outline of the special challenges mental models produce for the philosophy of mind (sec. 8).

1. The decline of behaviorism and the cognitive turn

Today mental models play a key role in psychology and cognitive science, but that was not always the case. Mental models basically ran through three different phases. In the early years of scientific psychology, phenomena were described that we nowadays interpret as involving the use of mental models, but they were described differently as the concept then did not exist. With the cognitive turn (in the 1950) the conception developed very quickly and soon became one of the central notions of cognitive psychology and cognitive science. In a third phase, the concept of mental models also appeared in the cognitive neurosciences, where researchers are now searching for the neural correlates of mental models.

In the first part of the last century there was no room for mental models in psychology. In his famous 1913 paper, Watson emphasized the study of observable behavior, rejecting introspection—the direct observation of one's own inner life—and theories of the (un-)conscious as unscientific approaches to psychology (Watson 1913). Following Watson's behaviorist proposal, mentalistic descriptions were banned altogether from the psychological vocabulary in favor of objective descriptions of behavior in dependence of stimuli. In psychological experiments, the stimuli had to be controlled systematically and the ensuing behavior objectively measured. The method's aim was to describe human (and animal) behavior as systematic regularities between input (stimuli) and output (behavior). The cognitive system itself (the human brain) was viewed as a black box, the internal

states of which are not amenable to scientific description or explanation. Therefore, the major concern of behaviorists was conditioning (the learning from stimulus-response combinations). They attempted to describe behavior of every kind, even as complex as linguistic behavior, entirely in terms of stimulus-response patterns (cf. Skinner 1938, 1974).

During the 1950s, more and more psychologists began to challenge the behaviorist dogma of the cognitive system as a black box. A crucial step in this development was a series of experiments with rats, described by Tolman in 1948. Though Tolman started from behaviorist premises, his results turned out to be unexplainable without a concept of mental representation. In one of the experiments, rats were trained to follow a path through a complex maze in order to reach a food box. After the rats had performed perfectly (chosen the shortest way to reach the goal), the trained path was blocked and the rats had to select another path from a variety of alternatives. Astonishingly, most of the rats found a path that was close to the most direct connection to the food box, whereas not a single rat erroneously tried to follow the original path on which they had been trained. On the basis of these results, Tolman argued that the rats must have acquired an internal representation (which Tolman did not call a "model") of the labyrinth. Today there is a large body of evidence on how humans (and animals) explore their environment and mentally represent it. Moreover, we now believe that such mental representations of spatial environments are constructed even if we do not directly experience the environment when navigating through it, but also when we just hear or read about it. Innumerable studies in the field of text comprehension have shown that mental models are routinely and immediately activated during word and sentence comprehension. If individuals are asked to read texts, they regularly construct a mental model of the (possibly fictitious) environment while reading (e.g. Glenberg 1997, Zwaan et al. 2002). As these results illustrate, today most psychologists are convinced that more can be said about a cognitive system than just registering input-output regularities. Its internal states, whatever they are, can be described in terms of the *functions* they have for the whole system. In these terms, descriptions of the system's state can be given that are, on the one hand, much more informative and detailed than a behaviorist would be willing to grant, but that are, on the other hand, entirely independent of the concrete implementation. Progress in the theory of computability was a major source for this new point of view (see next section).

The "cognitive turn", i.e. the switch from behaviorism to cognitive psychology, can be dated to the appearance of Ulric Neisser's 1967 book *Cognitive Psychology*, which showed the application of the new method to various areas in psychology. Soon, a new field called cognitive science arose from the combination of methods from various disciplines. The leading method,

Table 1
The description of a chocolate vendor automaton with initial state 1

State	Input	Output	Following State
1	50 Cent	none	2
1	1 Euro	chocolate	1
2	50 Cent	chocolate	1
2	1 Euro	chocolate, 50 Cent	1

however, was imported from artificial intelligence research. As machines are programmed to solve problems for which humans are said to require intelligence, humans themselves are viewed as such problem-solving systems. The rationale of this comparison is the fact that, with reference to analogous tasks, the states in both kinds of systems are functionally equivalent. Cognitive science today tries to combine the experimental methods of cognitive psychology with the computational methods of artificial intelligence in order to gain more insight into the functioning of the human mind.

In regard of the mentioned equivalence, the major premise of cognitive science is the "Physical Symbol System Hypothesis" (Newell & Simon 1976). It states that the human brain is essentially a physical symbol system. This implicates that cognition (what the human brain does) can be exhaustively described by computational methods because a symbol system does nothing but computation. For this reason, cognitive science is founded on conceptions of computational theory. It will be helpful to introduce some of these conceptions.

2. The computational view of mind and the levels of description

The main concept of automaton theory is that of an abstract automaton. An automaton is defined by states, in which the automaton can be, inputs, outputs, a function mapping every state and input to a subsequent state and output, and an initial state. For simple automata, this can be written in a table—see, e.g., the description of a chocolate vendor in table 1. The possible states of an automaton are hence defined solely by the function mapping input and actual automaton state to output and subsequent automaton state. Therefore, internal states of an automaton can be described *functionally* (in terms of their "functional roles"). The view that cognitive

systems are automata and hence their internal [1] states can be described functionally, is called functionalism (Fodor 1968, Putnam 2000).

In mathematics, the intuitive notion of an effectively calculable function has found several formalizations. One such famous formalization involves the Universal Turing machine, introduced by Alan Turing 1936. This machine is a virtual automaton and every function it can compute is called a Turing-machine computable function. Other formalizations comprise lambda-definable functions (Church 1932, Kleene 1935) and recursive functions (Gödel 1934, Herbrand 1932). It can be proved that all these formal analyses of the intuitive notion of an effectively calculable function are equivalent. This result gave rise to the hypothesis now known as the Church-Turing Thesis: Every function that is effectively calculable is Turing machine computable.

The states of an automaton can be described in terms of functions. The functions themselves can be viewed as effectively calculable in the sense of the Universal Turing machine. Hence, every automaton can be modeled by a Universal Turing machine. However, the Universal Turing machine is equivalent to other ways of modeling (lambda calculus, recursive functions, etc.). Therefore, the *functional* description of an automaton is just as good (informative) as an *algorithmic* description: it leads to a full description of an automaton in the sense defined above. There are many equivalent ways to express the functions. Moreover, every algorithm can be implemented in various ways. These considerations led to the characterization of three levels of description (Marr 1982): the implementation level, the algorithmic level, and the computational (functional) level. An automaton can be described on all three levels. However, the implementation level description does not offer essentially new information compared to the functional level; on the contrary, details about algorithms and their implementation are not interesting since there are many possible ways of implementing one and the same (functionally described) automaton.

Since the Universal Turing machine is a symbol processing machine, computation can be identified with symbol processing. In cognitive science, cognition is characterized as a form of computation, where internal states plus input are mapped to internal states plus output. It follows that every form of cognition can be done by a physical symbol processing machine. This hypothesis is exactly Newell and Simon's Physical Symbol System Hypothesis.

Adopting both the idea of different levels of description and the Physical Symbol System Hypothesis, human cognition can be fully described on a functional level. Because this level can be described regardless of how

[1] In the case of humans, internal states are often called mental states.

cognitive functions are implemented, human cognition can be implemented not only in the human brain, but in any system equivalent to the Universal Turing machine. Thus, also the functional description of cognition that stems from behavioral experiments can be implemented on a computer. The method of cognitive science is hence a combination of psychological methods and methods of artificial intelligence. For each theory (functional description) of cognitive phenomena there should be a possible implementation on a computer exhibiting the same phenomena. If both of these constraints (empirical accuracy and implementational possibility) are fulfilled, nothing of interest can be added by looking at the original implementation (the brain).

3. The doctrine of mental logic

In the early years of the 20th century, the developmental psychologist Jean Piaget had studied how children's ability to reason increases as they grow up. His research culminated in a theory of cognitive development stating that children of different ages are equipped with (or have access to) different inventories of inference rules as a basis for reasoning. Piaget's main assumption was that human reasoning relies on a *mental logic* consisting of formal inference rules. More than fifty years later, the computer metaphor of human cognition led to a renascence of the rule-based approach to reasoning. These theories, especially prominent in the 1970s, state that mental representations have the form of propositions, much like logical formulae. Reasoning is performed by applying syntactical rules to transform propositions, like in logical proofs (Johnson-Laird 1975, Osherson 1975, Braine 1978, Rips 1983). Because this view was dominant at the time, Johnson-Laird (1983) calls it the *doctrine of mental logic.*

The view that humans basically perform syntactical transformations of propositions fits very well with the program architecture known as rule-based systems. A rule based system has two memories, one for propositions (the declarative memory) and one for syntactical rules (the procedural memory). If the system is given a set of new propositions it is able to select and apply rules from the procedural memory. It will thereby generate new propositions (i.e. new information). This process can, if necessary, include propositions from the declarative memory, for example axioms and background knowledge. Therefore, theories of mental logic are very easy to program and they still are popular among cognitive scientists. However, it is clear that humans, in many contexts, do not reason logically sound. Especially if the context is poor and the reasoning task is abstract, humans fail to generate correct conclusions. This fact is usually explained by limi-

tations of the working memory that must hold all the propositions needed for a certain inference. The load on working memory, and consequently the difficulty of a task, increases with the number of rules to be applied. Nevertheless, if the rules in the procedural memory are abstract logical rules, the differences between reasoning with abstract and concrete material find no straightforward explanation. A further question is, how these logical rules are learned and whether it is plausible to assume people to have full abstract logical competence in that way. People with no background in psychology often report that they do not use logical derivations but rather construct—before their mind's eye—an integrated representation of the information given in the premises and then "read off" new information, not explicitly given in the premises. This is the fundament of the theory of mental models.

4. Mental model theory

How do humans draw inferences? In contrast with the mental logic doctrine, a layperson will quickly come up with the sensible idea that the content of all premises must be integrated into one 'picture.' Similarly, psychologists have conjectured that people integrate the information from the premises into a single mental representation. In 1943, for instance, Kenneth Craik claimed that the mind constructs "small-scale models" to anticipate events:

> If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it. (Craik 1943, 61)

Craik's idea is the germ of what we today know as mental model theory. This theory was first expounded by Philip Johnson-Laird in an article titled "Mental models in cognitive science" (Johnson-Laird 1980) and, in full detail, in his book *Mental models: towards a cognitive science of language, inference and consciousness* (Johnson-Laird 1983).[2] The main purpose of this work was the development of a theory of human thinking and reasoning that goes along with a critique of the mentioned doctrine of mental

[2] The editorial board of *Cognitive Science*, where Johnson-Laird's 1980 article appeared, has recently identified several classic articles from the journal from the last couple of decades. The members of the committee rated Johnson-Laird's article among the top ten because of its impact, innovation, and importance in promoting theoretical development in the field of cognitive science.

logic in cognitive science. Johnson-Laird showed that human thinking and reasoning can be modeled without propositional representations but, at the same time, saving the advantages of describing mental processes in computational terms. He argues that mental models rather than formal logic underlie syllogistic inferences, e.g., "All men are animals, the professor is a man, therefore the professor is an animal." The article was the first to present a case for mental models as a computational theory of human thought. Johnson-Laird extensively argued that different levels of description must be distinguished in order to describe cognitive processes. Below the behavioral level lies the computational level, and Johnson-Laird argued that a reasoning theory must be translated into a computer language in order to be executable on a computer. The researcher must specify the underlying representational format and the procedures that generate and manipulate this representation. However this description must be in *functional* terms, rather then in terms of the bits and bytes that move around in the computer hardware. Johnson-Laird thus endorses what we have called the independence of functional level and implementation level. This is expressed, e.g., in the following quotation:

> We should not worry about the particular computer and its machine code, since the program could be executed on some very different machines, and we do not want to make a different characterization for all these different sorts of computer. (Johnson-Laird 1980, 100)

According to mental model theory, human reasoning relies on the construction of integrated mental representations of the information that is given in the reasoning problem's premises. These integrated representations are the mental models. A mental model is a mental representation that captures what is common to all the different ways in which the premises can be interpreted. It represents in "small scale" how "reality" could be—according to what is stated in the premises of a reasoning problem. Mental models, though, must not be confused with images. A mental model often forms the basis of one or more visual images, but some of them represent situations that cannot be visualized (Johnson-Laird 1998). Instead, mental models are often likened to diagrams since, as with diagrams, their structure is analogous to the structure of the states of affairs they represent. From the processing view, the model theory distinguishes between three different operations. In the *construction phase*, reasoners construct the mental model that reflects the information from the premises. In the *inspection phase*, this model is inspected to find new information that is not explicitly given in the premises. In most variants of the model theory, the inspection process is conceptualized as a spatial focus that scans the model to find new information not given in the premises (Bara et al. 2001, Ragni et al. 2005, Schlieder & Berendt 1998). In the *variation phase*, rea-

soners try to construct alternative models from the premises that refute the putative conclusion. If no such model is found, the putative conclusion is considered true.

The theory's central idea is that of an analogy of a mental model's structure with the structure of the situation modeled. In this respect, mental models, much like the pictures of Wittgenstein's 1922 "picture theory", represent a certain situation conceived in just one possible way. Sometimes a model captures what is common to all of the different ways in which the possibility may occur. Then it is a perfect basis for reasoning. However, sometimes—and in fact most frequently—reasoners are unable to survey the entire set of possibilities and thus focus on a subset of possible models—often just a single model—which leads to incorrect conclusions and illogical decisions. It is interesting that humans have preferences if a problem has multiple solutions and that most people agree in their preferences. For a certain task, we tend to construct almost the same single model—the preferred mental model—and to ignore others (Knauff et al. 1998).

The crucial difference with theories of mental logics is that no knowledge of logical rules must be presupposed. The reasoner constructs and manipulates mental models not according to abstract logical rules but according to the world which she represents. After having integrated all the information of the premises in one (or more) consistent models, the conclusion can be directly "seen" in the model (and eventually compared with conclusions from other models). In this way, logically sound reasoning "emerges" from the format of representation. Failure of sound reasoning can be explained, as sketched above, from the fact that not all relevant models are constructed for many problems.

As illustrated in this book, cognitive psychologists have explored mental models from very different points of view and carried out an extensive research program on how models engender thoughts and inferences. Since the understanding of (linguistically presented) premises involves text comprehension, the theory has been extended to provide a psycho-linguistic approach to semantic processing. In the field of education, the role of mental model construction in learning has been explored. The question whether mental model theory also contributes to the understanding of (visual) perception is currently discussed in cognitive psychology and the philosophy of perception. These are just a few examples for the immense influence of mental model theory across the borders of academic disciplines. Today, this research effort is much more successful than the classical rule based approaches of reasoning.

5. Mental models and the brain

In the last years, the position of the implementation-independency has been graded down because most cognitive scientists now believe in the assumption that the understanding of brain-events can provide insight into the computations they implement (cf. Gazzaniga et al. 2002, Johnson-Laird 1995). That is what we call the third phase of mental models research. Still, behavioral methods are the *via regia* to understand human thinking and reasoning with mental models. Today however, neuroscientific research is adding important information about the characteristics of mental models. Studies with brain-injured patients gave us an initial idea about which areas of the brain are involved in thinking and reasoning and the availability of modern brain imaging methods currently contributes enormously to our understanding of human thought and behavior.

Researchers with a background in mental model theory differ in some respects from other fields of cognitive neuroscience. They are still cognitive psychologists with the goal to understand human experience and behavior and the intervenient computational processes. They are not so much interested in what Uttal (2001) called the *new phrenology*, namely the localization of cognitive processes including all the reductionistic implications. Instead, they treat changes in cortical blood flow as a dependent variable, much as response times or error rates. The background of this new turn in mental models research is that the mentioned "independence of computational level" hypothesis makes some questionable assumptions, after all. In particular, the supposition that each function computable by a Turing Machine can be computed on all Turing-equivalent machines is not unqualifiedly true (Giunti 1997, Goel 1995). Though it is true that computational processes can be realized in *many* different systems, it is not true that they can be realized in *all* Turing-equivalent machines. The assumption of universal realizability thus appears to be unwarranted (Goel 2004). A second reason for the new interests of mental model researchers is that localization and dissociation can help to understand the cognitive processes themselves. As Goel (2004) puts it, Gall & Spurzheim (1810-1819) was basically right and Lashley (1929) wrong about the organization of the brain. Not all neural computations can be realized in all brain areas. We know that there are highly specific brain regions dedicated to specific computations. For instance, there are brain areas that exclusively process information in a verbal format, whereas other areas only respond to incoming information in a visuospatial format. For the testing of hypotheses it is essential that these cortical systems can be identified with rule-based and model-based reasoning processes. Language-related brain areas are often identified with rule-based theories of thinking, whereas activity in visuospatial brain ar-

eas during thinking is typically identified with mental models. A similar distinction is related to the two brain hemispheres, where the right brain is related to the processing of mental models and the left brain to more language-based abstract inference processes. Such localization can help us to test different cognitive theories, since different theories—namely mental logic and mental model theories of reasoning—make different predictions concerning the involved brain areas.

6. Mental representation as a philosophical problem

Mental models are mental representations of a certain type. The main problem in the philosophy of mental representation is to characterize the relation between a mental representation and the represented object. Naively speaking, a mental representation is an entity that 'stands for' another—the represented object—, but here 'stands for' is just a metaphoric place-holder for 'represents', thus requires further explanation. Obvious features of the representation relation can be isolated. First, it is an asymmetric relation (if X represents Y, then Y usually does not represent X); and, second, there are cases of misrepresentation, where, e.g., a cognitive system represents a horse as a cow. In the recent literature on representation there are three main types of representation theories tackling these problems: causal theories, similarity theories, and functional theories.

The basic idea of the first type of theories is that some mental entity represents another (non-mental or mental) entity because the first is caused by the second. In this way, the asymmetry of the relation is straightforwardly explained. However, the problem of misrepresentation is much harder to deal with. As sometimes Xs cause Y-representations (misrepresentation), Y-representations should have a disjunctive meaning "X or Y" according to the causal approach. To avoid this disjunction problem, Fodor (1987) introduces a nomic relation between the Y-caused Y-representations and the X-caused Y-representations such that X-caused Y-representations can occur only when there are Y-caused Y-representations, but not vice versa.

Similarity theories are based on the assumption that representations are similar to what they represent. A general problem for these theories obviously is the explanation of asymmetry.[3] However, the main problem seems to be to characterize the kind of similarity. Indeed, many different kinds have been proposed (cf. Cummins 1989). Nowadays, the most attractive similarity relation seems to be isomorphism, proposed in both philosophy (e.g. Cummins 1996, French 2002) and in psychology (especially for mental

[3] Although there are some similarity relations that are non-symmetric.

models; e.g. Palmer 1978, Gurr 1998). Still, it remains an open question what misrepresentation is: A Y-representation is not similar to an X in the relevant sense (otherwise it would be an X-representation), but still can be erroneously taken to represent X.

The last group of theories imports the notion of function from biology. Mental representations can be characterized by the functions they have and eventually fulfill, just like organs or traits of an organism are characterized by their functions. A heart is an entity that can be fully characterized by its function of pumping blood (independent of whether it actually fulfills this function or not). In the very same way, a mental representation is an entity that has the function to represent (e.g. Dretske 1994, Millikan 1984, 1989). The function of a horse-representation, for example, is to stand for the horse within the functional architecture of a cognitive system. Because of this function in the system it leads to certain states of the system concerning the horse. A cow-representation, on the other hand, would be, in an obvious sense, dysfunctional in this context. It does not lead to states concerning the horse (but to states concerning a cow that is not there) and thus does not (indeed cannot) fulfill the function to stand for a horse.

7. Intentionality and its naturalization

The mentioned theories try to come to grips with asymmetry and misrepresentation. What about the original feature that a mental representation 'stands for' the object it represents? This property traditionally is an intentional property because it is possible that a mental representation stands for something non-existent (as the cow-representation did in the previous example). Being a relation to something possibly non-existent, arguably, is the mark of the intentional. More exactly, intentional states are goal-directed, but these goals need not to exist. This non-existence implies that the states are open-ended in the sense that they can have or not have a relatum (see, e.g. Chisholm 1957, 170). Representations are intentional and open-ended states in this sense: they are directed toward the entities they stand for and these entities need not exist even if the representations do.

Obviously, the properties of a mental representation's reference and intentionality are in urgent need of further explanation, but philosophers are deeply divided about what form that explanation should take. One large group thinks that such explanation should proceed in terms of relations of agents with their environment that are more readily accessible to scientific treatment and direct observation. This project is often described as "naturalizing the mind" (see, e.g. Loewer 1997). Others think that this project is deeply misguided (see, e.g. Putnam 1982). The latter attitude,

well-founded as it may be, seems unhelpful in view of the fact that the cognitive sciences do already investigate the mind with techniques that are both very successful and regarded as scientifically acceptable.

One way to summarize the qualms about the naturalization project is as follows. The project is successful where it underpins empirical research on the mental as we see it done today, because it rectifies suspicious philosophical vocabulary into scientifically acceptable terminology. However, it fails where it aims to explain mental states in scientific terms by eliminating the intentional vocabulary because the latter either will be tacitly reintroduced to make naturalized descriptions applicable (in the case of physical descriptions) or is entrenched in the scientific terminology from the outset (in the case of biological descriptions).

A good starting point to illustrate this fact is, again, the "Physical Symbol System Hypothesis." In Simon's words, the hypothesis states "that a physical symbol system [...] has the necessary and sufficient means for general intelligent action" (Simon 1996, 23). As we saw, the hypothesis is interpreted as saying that humans think and represent by manipulating symbols and as such it has been a fruitful research hypothesis in cognitive science. The ensuing hypothesis that humans are physical symbol systems is a helpful tool to make several of their mental activities scientifically accessible in a new and fruitful way. Intentional language is imprecise and thus, in comparison with, say, an algorithm for symbol manipulation, less suited for describing such activities exactly. However, naturalization is at work here only in the sense that the mental is made scientifically accessible by means of new tools. Only when physical symbol systems, and thus humans, are interpreted as mere physical systems a more serious naturalization project is initiated. Clearly, Simon himself aims in this direction. He introduces symbols as physical patterns with the clear intention of interpreting physical symbol systems as purely physical systems. The proposal is interesting, but ultimately doomed to fail—or so the skeptic will argue. Simon writes: "Symbol structures can, and commonly do, serve as internal representations (e.g., 'mental images') of the environment to which the symbol system is seeking to adapt." (Simon 1996, 22) It is here that the skeptic will claim an ill-reflected re-introduction of intentional vocabulary ("serve," "seek," "adapt") into a context that pretends to be pure physics.

In cognitive science, the reliance on the program advocated by Simons is history. As is reasonable, these sciences today make unrestricted use of biological vocabulary. Do attempts to explain mental representation in this vocabulary count as naturalizations? A key notion of biological explanations is the one of function. Indeed, we can understand the above-mentioned functionalist theory of mental representation as a naturalization project, if the mental representations having (and fulfilling or failing to fulfill) certain functions are *identified* with organic states having these functions. Here,

it is the notion of function where the skeptic suspects a vicious circle. To explain intentional relations, naturalizers utilize a teleological notion from biology, which is itself in need of explanation. Philosophers of biology have long quarreled about the appropriate characterization of biological functions (see, e.g. Sober 2000). Attempts at a coherent and satisfying characterization differ in many important details, but nowadays they mainly appeal to natural selection as an explanation of a biological subsystem's function for the including system. But it is specifically selection in animate systems, thus in systems that exhibit biological activity, that is exploited for philosophical accounts of biological functions, as, e.g., when the chance of reproduction under selective pressure is characterized as the chance to survive and/or sexually reproduce. So, those who employ functional notions to describe an entity tacitly refer to intentional notions via the goal-directedness of the very organism of which the functional or dysfunctional element is an organic part or state. E.g., explaining a frog's mental representation of a fly in terms of a state that functions to aid the frog in catching its food is to utilize a notion that itself makes tacit appeal to the frog's activity of trying to eat, because that is what that state, if it functions, contributes to. In turn, the fact that an equal state contributed, via contribution to successful eating behavior, to the survival and sexual reproduction of the frog's ancestors, explains the state's existence in the frog, but this evolutionary explanation of the functional element presupposes a goal-directed activity of the frog and its ancestors.

Without scientific explanation the intentional phenomena remain mysterious and it seems that such explanation must take the course of naturalization. So far, however, only weak naturalization—making the mental scientifically accessible while consciously preserving its description in the intentional vocabulary—is a successful project. And it appears that the sciences don't need more naturalization.

8. Mental models and the philosophy of mind

The psychological insight that humans use mental models in many cognitive processes gives several issues in the philosophical debate a new twist. The status of such models as mental entities stands and falls with the one of mental representations in general. However, for those philosophers who want to show that mental representations, as entities in their own right and with their own distinctive features, do ultimately not exist, mental models raise the bar. Mental models have initially been proposed as special representations that explain how humans reason, so the philosopher denying

the existence of mental representations faces the challenge: "Explain how people think!"—but without mental models (see Johnson-Laird 1996, 90).

Mental models offer still more challenges. After all, cognitive psychologists distinguish them from other types of representation through their specific functions (Johnson-Laird 1996, 91). Models are assumed to represent classes of situations as opposed to images representing single situations (Johnson-Laird 1996, 120). This idea is based on the premise that models contain abstract elements, which is in turn based on the assumption that they have the *function* of aiding the execution or evaluation of syllogistic reasoning. Thus, it is essential to a philosophical understanding of mental models to *functionally* differentiate them from other types of representation.

This differentiation involves several aspects: Firstly, it has to be described in what respect the relation between a model and the represented situation differs from the relations between other forms of representation and the objects they represent. Secondly, the question of what mental models represent at all has to be answered. Thirdly, the status of the models itself within the cognitive system has to be contrasted with the status of other forms of mental representation. This differentiation involves not only a description of what a model is for the reasoner, but also an answer to the question what kind of entities mental models are ontologically.

References

Bara, B., Bucciarelli, M. & Lombardo, V. (2001), 'Model theory of deduction: A unified computational approach', *Cognitive Science* **25**, 839–901.

Braine, M. D. S. (1978), 'On the relation between the natural logic of reasoning and standard logic', *Psychological Review* **85**, 1–21.

Chisholm, R. (1957), *Perceiving: A Philosophical Study*, Cornell University Press, Ithaca.

Church, A. (1932), 'A set of postulates for the foundation of logic', *Annals of Mathematics, second series* **33**, 346–366.

Craik, K. (1943), *The Nature of Explanation*, Cambridge University Press, Cambridge.

Cummins, R. (1989), *Meaning and Mental Representation*, The MIT Press, Cambridge, MA, London.

Cummins, R. (1996), *Representations, Targets, and Attitudes*, The MIT Press, Cambridge, MA, London.

Dretske, F. (1994), Misinterpretation, *in* S. Stich, ed., 'Mental Representation. A Reader', Blackwell, Cambridge, MA, Oxford, pp. 157–173.

Fodor, J. (1987), *Psychosemantics*, The MIT Press, Cambridge, MA, London.

Fodor, J. A. (1968), *Psychological Explanation*, Random House, New York.

French, S. (2002), 'A model-theoretic account of representation', *Proceedings of the PSA* (Supplement).

Gall, F. J. & Spurzheim, J. (1810-1819), *Anatomie et Physiologie du Système Nerveux en Général et du Cerveau en Particulier*, F. Schoell, Paris.

Gazzaniga, M. S., Ivry, R. & Mangung, G. R. (2002), *Cognitive Neuroscience: The Biology of the Mind*, 2nd edn, W.W. Norten.

Giunti, M. (1997), *Computation, Dynamics and Cognition*, Oxford University Press, New York.

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

Gödel, K. (1934), 'On undecidable propositions of formal mathematical systems', *Lecture notes taken by Kleene and Rosser at the Institute for Advanced Study* (Reprinted in M. Davis, ed., 1965, 'The Undecidable', New York, Raven).

Goel, V. (1995), *Sketches of Thought*, MIT Press, Cambridge, MA.

Goel, V. (2004), Can there be a cognitive neuroscience of central cognitive systems?, *in* D. Johnson & C. Erneling, eds, 'Mind as a Scientific Object: Between Brain & Culture', Oxford University Press, Oxford.

Gurr, C. A. (1998), On the isomorphism, or lack of it, of representations, *in* K. Marriott & B. Meyer, eds, 'Visual Language Theory', Springer, New York, Berlin, Heidelberg.

Herbrand, J. (1932), 'Sur la non-contradiction de l'arithmetique', *Journal für die reine und angewandte Mathematik* **166**, 1–8.

Johnson-Laird, P. (1996), Images, models, and propositional representations, *in* M. de Vega, M. Intons-Peterson, P. Johnson-Laird, M. Denis & M. Marschark, eds, 'Models of Visuospatial Cognition', Oxford University Press, New York, pp. 90–127.

Johnson-Laird, P. N. (1975), Models of deduction, *in* R. J. Falmagne, ed., 'Reasoning: Representation and Process in Children and Adults', Erlbaum, Hillsdale.

Johnson-Laird, P. N. (1980), 'Mental models in cognitive science', *Cognitive Science* **4**, 71–115.

Johnson-Laird, P. N. (1983), *Mental Models*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P. N. (1995), Mental models, deductive reasoning, and the brain, *in* S. Gazzaniga, ed., 'The Cognitive Neurosciences', MIT Press, Cambridge, MA, pp. 999–1008.

Johnson-Laird, P. N. (1998), Imagery, visualization, and thinking, *in* J. Hochberg, ed., 'Perception and Cognition at Century's End', Academic Press, San Diego, CA, pp. 441–467.

Kleene, S. C. (1935), 'A theory of positive integers in formal logic', *American Journal of Mathematics* **57**, 153–173, 219–244.

Knauff, M., Rauh, R., Schlieder, C. & Strube, G. (1998), Mental models in spatial reasoning, *in* C. Freksa, C. Habel & K. F. Wender, eds, 'Spatial Cognition—An Interdisciplinary Approach to Representing and Processing Spatial Knowledge', Springer, Berlin, pp. 267–291.

Lashley, K. S. (1929), *Brain Mechanisms and Intelligence: A Quantitative Study*

of Injuries to the Brain, University of Chicago Press, Chicago.

Loewer, B. (1997), A guide to naturalizing semantics, *in* B. Hale & C. Wright, eds, 'A Companion to the Philosophy of Language', Blackwell, Oxford, pp. 108–126.

Marr, D. (1982), *Vision: A Computational Investigation in the Human Representation of Visual Information*, Freeman, San Francisco.

Millikan, R. G. (1984), *Language, Thought, and Other Biological Categories*, The MIT Press, Cambridge, MA, London.

Millikan, R. G. (1989), 'Biosemantics', *The Journal of Philosophy* **86**(6), 281–297.

Neisser, U. (1967), *Cognitive Psychology*, Appleton-Century-Crofts, New York.

Newell, A. & Simon, H. (1976), 'Computer science as empirical inquiry: Symbols and search', *Communications of the Association for Computing Machinery* **19**, 113–126.

Osherson, D. N. (1975), Logic and models of logical thinking, *in* R. J. Falmagne, ed., 'Reasoning: Representation and Process in Children and Adults', Erlbaum, Hillsdale.

Palmer, S. (1978), Fundamental aspects of cognitive representation, *in* E. Rosch & B. L. Lloyd, eds, 'Cognition and categorization', Erlbaum, Hillsdale, NJ, pp. 259–302.

Putnam, H. (1982), 'Why reason can't be naturalized?', *Synthese* **52**, 3–23.

Putnam, H. (2000), *The Threefold Cord*, Columbia University Press.

Ragni, M., Knauff, M. & Nebel, B. (2005), A computational model of human reasoning with spatial relations, *in* 'Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Society', Erlbaum, Mahwah, NJ.

Rips, L. J. (1983), 'Cognitive processes in propositional reasoning', *Psychological Review* **90**(1), 38–71.

Schlieder, C. & Berendt, B. (1998), Mental model construction in spatial reasoning: A comparison of two computational theories, *in* U. Schmid, J. F. Krems & F. Wysotzki, eds, 'Mind Modelling: A Cognitive Science Approach to Reasoning, Learning, and Discovery', Pabst Science Publishers, Lengerich, pp. 133–162.

Simon, H. (1996), *The Sciences of the Artificial*, 3rd edn, MIT Press, Cambridge, MA.

Skinner, B. F. (1938), *Behavior of Organisms*, Appleton-Century-Crofts, Inc., New York.

Skinner, B. F. (1974), *About behaviorism*, Knopf, New York.

Sober, E. (2000), *Philosophy of Biology*, Westview Press, Boulder, CO.

Tolman, E. C. (1948), 'Cognitive maps in rats and men', *Psychological Review* **55**, 189–208.

Turing, A. M. (1936), 'On computable numbers, with an application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society* **42**(2), 230–265.

Uttal, W. R. (2001), *The New Phrenology*, MIT Press, Cambridge, MA.

Watson, J. B. (1913), 'Psychology as a behaviorist views it', *Psychological Review* **20**, 158–177.

Wittgenstein, L. (1922), *Tractatus Logico-Philosophicus*, Routledge & Kegan Paul, London.

Zwaan, R. A., Stanfield, R. A. & Yaxley, R. H. (2002), 'Do language comprehenders routinely represent the shapes of objects?', *Psychological Science* **13**, 168–171.

Part I

Cognitive Psychology

This Page is Intentionally Left Blank

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Introduction: Cognitive Psychology

Competence, difficulty, content—these three keywords stand for the subjects central to the study of human thinking and reasoning. They mirror three research questions: By which mechanisms can individuals reason? What factors cause reasoning difficulty? And: How do content and background knowledge affect reasoning performance? In the last decades, reasoning research made much progress in answering these questions. In specific cases, we think by applying mental rules, which are similar to rules in computer programs. In most of the cases, however, we reason by constructing, inspecting, and manipulating mental models. These models and the processes that manipulate them are the basis of our competence to reason. In general, it is believed that humans have the competence to perform such inferences error-free. Errors do occur, however, because reasoning performance is limited by capacities of the cognitive system, misunderstanding of the premises, ambiguity of problems, and motivational factors. Moreover, background knowledge can significantly influence our reasoning performance. This influence can either be facilitation or an impedance of the reasoning process. Technically speaking, the abstract (logical) truth value of an inference can be the same as the truth value of our prior knowledge—in this case the inference is supported. Or, the formal truth value conflicts with the truth value of the prior knowledge—then the inference is more difficult, which means it results in more errors or takes significantly longer.

The first three chapters of this book are all concerned with the mechanisms of reasoning, the causes for errors, or with the connection between reasoning and prior knowledge. **Johnson-Laird** himself uses the model theory to explain how individuals reason with sentential connectives, such as "if", "or", and "and" and why we commit errors in such task. He shows that certain inferences yield systematic fallacies and that these fallacies can be explained perfectly by the use of mental models. The chapter explains the models theory's predictions and reports some studies corroborating the

occurrence of these "illusory" inferences. His research is impressive because no other theory is able to explain the experimental findings.

The chapter by **Vandierendonck, Dierckx, and Van der Beken** is concerned with the connection between mental models and background knowledge. Vandierendonck explains this connection in the field of relational reasoning and shows that this kind of inference is based on an interaction of knowledge represented in semantic and episodic long-term memory on the one hand and temporary information maintained in working memory on the other hand. His account is very plausible and fits nicely with many experimental findings: Reasoners have a preference for visuo-spatial representations, believability affects reasoning performance, and reasoning with transitive and intransitive relations is related to different types of prior knowledge. Overall, the chapter shows that reasoning (with relations) is based on a tight interplay of knowledge representations in long-term memory and temporary models in working memory.

The chapter by **Seel** revolves around the function of mental models in learning. The author is a pedagogue and thus interested in the potentials of mental models to facilitate learning. For him, learning situations require the construction and manipulation of mental models. His main argument is that models support the simplification and visualization of the learning materials. The chapter reports on two empirical investigations that emphasize the facilitating effects on models in multimedia learning and discovery learning.

All three chapters of this part of the book indicate the likely direction of future empirical research. Firstly, behavioural experiments will continue to be the *via regia* to study human thinking and reasoning by means of mental models. They will continue to be the most helpful means to understand the nature of human reasoning, in particular if they are—as in the next chapters—combined with methods from cognitive neuroscience. Secondly, mental models researchers will continue to suggest modifications and refinements to explain new experimental findings. And finally: The theory of mental models will find its way into applications. The use of mental models in learning research is one example. Many other examples come from computer science, especially from artificial intelligence, where the orthodox view that logic representations together with forms of logical inference are sufficient to exhibit intelligent behavior is complemented—and even rejected—by representations in the form of mental models.

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Mental Models, Sentential Reasoning, and Illusory Inferences

P.N. Johnson-Laird [1]

Department of Psychology, Princeton University [2]

Abstract
This chapter describes how individuals reason with sentential connectives, such as "if," "or," and "and." They do not have a "truth functional" semantics for these connectives, but rather they construct models of the possibilities compatible with sentences in which the connectives occur. Human working memory has a limited processing capacity, and so individuals aim to construct only a single model at a time, and to represent only those clauses in the premises that hold in each possibility. One unexpected consequence of the theory emerged from its computer implementation. Certain inferences should yield systematic fallacies if reasoners use mental models. The chapter explains this prediction and reports some studies corroborating the occurrence of these "illusory" inferences. No one has yet devised an account of them on the basis of another theory.

Suppose that you are carrying out a test of system and you know that if the test is to continue then the reactivity of the system must not have reached the critical level. You then observe that the reactivity has reached the critical level. What should you do? It seems obvious that you should stop the test. The engineers in charge at Chernobyl were in this position, but they continued the test (see Medvedev 1990). Why they continued is puzzling, because the test was not only dangerous, but pointless. It led to

[1] This research was supported by a grant from the National Science Foundation (BCS-0076287) to study strategies in reasoning. For their helpful advice, I thank Ruth Byrne, Vittorio Girotto, Geoff Goodwin, Uri Hasson, Karl Christoph Klauer, Louis Lee, Markus Knauff, Walter Schroyens, André Vandierendonck, Clare Walsh, and Yingrui Yang.
[2] E-mail: phil@princeton.edu

the disaster. One possibility is that the engineers failed to make a valid inference of the form:

If A then not B.

B.

Therefore, not A.

where *A* stands for "the test is to continue" and *B* stands for "the reactivity has reached the critical level."

For several years, I have given groups of engineering students a similar problem with an abstract content, such as:

If there is a triangle on the board then there is a circle on the board.

There isn't a circle on the board.

What, if anything, follows?

Typically, more than half of them respond that nothing follows from these premises. In fact, the premises yield the conclusion:

There is not a triangle on the board.

This conclusion is *valid*: it must be true given that the premises are true. But, the inference is quite difficult to make. The engineers are not reluctant to make inferences, because with premises of this sort:

If there is a triangle on the board then there is a circle on the board.

There is a triangle on the board.

nearly all of them draw the valid conclusion:

There is a circle on the board.

People do make mistakes, and the difference in difficulty between the two previous inferences is one of the most robust effects in the psychology of reasoning (see, e.g. Evans et al. 1993). Yet, reasoners are not always wrong. Psychologists therefore need to explain both their logical ability and the cause of their mistakes.

My aim in this chapter is to describe the mental mechanisms underlying a major sort of reasoning, so-called "sentential reasoning", which is based on negation and sentential connectives, such as "if," "or," and "and." The account is a development from the theory of mental models (Johnson-Laird 1983, Johnson-Laird & Byrne 1991). The theory postulates that the mind constructs models of the world that it uses to reason. It constructs them from perception (Marr 1982), imagination (Metzler & Shepard 1982), knowledge (Gentner & Stevens 1983), and the comprehension of discourse (Stevenson 1993, Polk & Newell 1995, Oakhill & Garnham 1996, Garnham 2001). A crucial distinction between models and other sorts of proposed mental representation is that the structure of models corresponds to the structure of what they represent: individuals are represented by individual tokens, properties by properties of these tokens, and relations by relations among these tokens (see, e.g. Johnson-Laird 1983).

In reasoning, a key step is to establish a conclusion; its strength depends on whether any models of the premises refute it (Johnson-Laird & Byrne

1991). The theory therefore provides a unified account of reasoning about what is necessary, probable, or possible. A conclusion is *necessary* if it holds in all the models of the premises, it is *probable* if it holds in most models of the premises (Johnson-Laird et al. 1999), and it is *possible* if it holds in at least one model of the premises (Bell & Johnson-Laird 1998).

The model theory, as I refer to it, is based on a core principle that concerns the interpretation of connectives, and that gives rise to systematic fallacies. These fallacies can be so compelling that they have an illusory quality: it is hard to avoid succumbing to them even when you are on guard against them. You will understand the principle more easily if I outline elementary logic. Hence, the chapter begins with such an account. It then describes the interpretation of connectives in natural language, and illustrates the limitations of human working memory. These limitations lead to the fundamental principle of the model theory: Mental models are parsimonious. The chapter formulates the mechanisms that implement this principle in the construction of mental models, which it contrasts with the reasoning of superhuman entities with unlimited working memories. It reports some illustrative results of recent studies of the illusory inferences. These results corroborate the theory.

1. Logic and truth-functional connectives

Logic treats sentences as expressing propositions; in everyday life, however, the proposition that a sentence expresses almost always depends on its context. "I can hear you now"—an utterance all too common these days—expresses different propositions depending on who says it, to whom it is addressed, and the time and circumstances of the utterance. To keep matters simple, I will use sentences that depend as little as possible on their context, and, where feasible, I will adopt the fiction that sentences are propositions.

Logic is the science of valid inferences. It is not concerned with how people make such inferences. Logicians have formulated many different calculi for formalized languages. They can set up a calculus in two distinct ways (see, e.g. Jeffrey 1981). The first way is formal, concerning patterns of symbols, but not their interpretation. The sentential calculus concerns sentential connectives in their logical senses—a notion that I will explain soon. Its *formal* specification depends on rules of inference, such as:

A or B, but not both.
not-B
Therefore, A.

Table 1
The truth table for an inclusive disjunction

There is a circle on the board.	*There is a triangle.*	*There is circle on the board or a triangle or both.*
True	True	True
True	False	True
False	True	True
False	False	False

The variables, A and B, can have as values any declarative sentences whatsoever.

The second way to characterize a calculus is *semantic*. Consider an atomic sentence, i.e., one that contains neither negations nor connectives:

There is a circle on the board.

Let's suppose that it is false. A *compound* sentence is made from atoms by combining them with negation or sentential connectives. Here is a negative compound:

There is not a circle on the board.

This assertion is true because, as I just told you, the atom that it contains is false. Suppose that you also know another compound assertion, which is a disjunction of two atoms:

There is a triangle on the board or there is a circle, or both.

This disjunction is *inclusive*, because it allows that both atoms could be true. Hence, its meaning is compatible with three possibilities:

There is a triangle on the board and there is not a circle.

There is not a triangle on the board and there is a circle.

There is a triangle on the board and there is a circle.

You already know that there is not a circle, and so you can eliminate all but the first possibility. It follows that that there is a triangle. The formal rule above also allows you to make this inference, but here you have made it on a semantic basis. Hence, in principle, human reasoning could be based on formal procedures or semantic procedures, or both.

The meaning of the preceding disjunction can be laid out in the form of a truth table, which specifies the truth value of the disjunction for each of the four possible contingencies—the three possibilities in which it is true, and the remaining possibility in which it is false. Table 1 presents this truth table. Each row in the table shows a possible combination of the truth values of the two atoms, and the resulting truth value of their inclusive disjunction. For example, the first row is the possibility in which both atoms are true, and, as a result, the inclusive disjunction is true too.

Truth tables were invented by the great American logician, Charles Sanders Peirce (see, e.g. Berry 1952), though Wittgenstein (1922) is often wrongly credited with their invention.

A sentential connective has a "logical" meaning when its interpretation can be summarized in a truth table. The truth table shows how the truth value of a sentence containing the connective depends solely on the truth values of its constituent propositions. Once you know their truth values, you can work out the truth value of the sentence as a whole from the connective's truth table. Hence, an inclusive disjunction in its logical sense is true or false solely as a *function of* the truth values of the constituent propositions. As logicians say, a disjunction has a *truth-functional* meaning. This piece of jargon means: you feed in truth values of the constituent propositions, and the truth table for "or" gives an output of a truth value.

In logic, a general recipe exists for interpreting compound sentences. You replace each atom with its truth value—how you obtain such truth values is not part of the theory—and you progressively simplify the compound according to the interpretation of each connective, until you arrive at a final truth value for the sentence as a whole. This truth value depends only on the truth values of the atoms and the truth-functional interpretations of the connectives.

Here is an example of such an interpretation. Consider the compound assertion in which "or else" is an *exclusive* disjunction, i.e., only one of the two clauses it connects is true:

(A and not B) or else (C and D)

and assume that all the atoms are true: A, B, C, and D, are all true. The first step in the interpretation of the compound is to replace its atoms with their truth values:

(true and not true) or else (true and true)

The next steps simplify the expression according to the truth-functional meanings of negation and the connectives:

(true and false) or else (true and true) —according to the meaning of *not*

(false or else true) —according to the meaning of *and*

true —according to the meaning of *or else*

Hence, the compound assertion is true given the values of its atoms.

Logicians can use truth tables to determine whether or not an inference is valid: It is valid if its conclusion must be true given that its premises are true. One of the glories of twentieth century logic was Gödel's discovery that there are logics in which not all inferences that are valid in their semantic system can be proved using a consistent formal system (see, e.g. Boolos & Jeffrey 1989). (The reason for the stipulation that the system

is consistent is that an inconsistent system would allow any proposition including contradictions to be proved.) The logic of sentential connectives, however, has the happy property that all inferences that are valid on the basis of their truth-functional meanings are also provable in a consistent formal system, and vice versa.

2. The interpretation of connectives in natural language

The psychology of reasoning would be simpler if all connectives in natural language were truth functional. But, temporal connectives, such as "and then" or "before," are not truth functional. It is true that Bush declared war on terrorism, and that terrorists attacked the USA, but the following assertion is nevertheless false:

Bush declared war on terrorism and then terrorists attacked the USA.

The two events occurred in the opposite order.

In fact, the human interpretative system cannot be truth functional, not even in the case of logical interpretations. As the example in the previous section showed, a truth-functional interpretation starts and ends with truth values. It doesn't take into account what individual atoms mean, what they refer to, or any temporal, spatial, or other such relation between them: All it depends on are truth values. When you understand a sentence, however, you don't end up with its truth value. Indeed, you may never know its truth value, which depends on the relation between what it signifies and the state of the world. Comprehension starts with the construction of the meaning of a sentence; it recovers its referents, their properties, and any relations among them—a process that may depend on knowledge; and it ends with a representation of the possible situations to which the sentence refers. In short, it starts with meanings and ends with models. The moral is clear. No connectives in natural language are interpreted in a truth functional way (see Johnson-Laird & Byrne 2002, Byrne 2005).

Many uses of "if," "or," and "and" don't have a logical meaning. The connective *and* can be interpreted to mean *and then.* The following disjunction:

They played soccer or they played some game

seems innocuous. But, if you learn that the second atom is false, i.e.:

They didn't play any game

you would not infer the truth of the first atom:

They played soccer.

The formal rule I presented earlier would allow this inference to be made, but in real life you wouldn't make it. You know that soccer is a game, and so you interpret the disjunction to be compatible with only two possibili-

ties, and in both of them they played a game (see Johnson-Laird & Byrne 2002 for an account of such "modulations" of interpretation). Hence, the disjunction no longer has a logical meaning.

3. The limitations of working memory

A *mental* model represents a possibility, or, to be precise, the structure and content of the model capture what is common to the different ways in which the possibility could occur—a construal that I owe to a logician, the late Jon Barwise (1993). When you are forced to try to hold in mind several models of possibilities, the task is difficult. To experience this phenomenon of "memory overload" for yourself, try the following problem:

June is in Wales or Charles is in Scotland, or both.
Charles is in Scotland or Kate is in Ireland, or both.
What, if anything, follows?

The disjunctions are inclusive, and so each premise is consistent with three possibilities. The problem, of course, is to combine the two sets of possibilities. In fact, they yield five possibilities, which support the valid conclusion:

June is in Wales and Kate is in Ireland, or Charles is in Scotland, or both.

Five possibilities are too many to hold in mind at the same time, and so, as the theory predicts, this inference is hard. My colleagues and I tested a sample of the general population in an experiment, and only 6% of them drew a valid conclusion (Johnson-Laird et al. 1992). The experiment also examined similar inferences based on exclusive disjunctions:

June is in Wales or Charles is in Scotland, but not both.
Charles is in Scotland or Kate is in Ireland, but not both.
What, if anything, follows?

These premises are compatible with only two possibilities, and they yield the conclusion:

Either June is in Wales and Kate is in Ireland or else Charles is in Scotland.

The problem was easier: 21% of the participants drew this conclusion or an equivalent to it.

Most people go wrong with both sorts of inference, and so you might wonder what conclusions they draw. If they are trying to construct mental models of the various possibilities, then there are two obvious predictions. The first is that if they grasp that there's more than one possibility but are unable to discern what holds over all of them, then they should respond that there's no valid conclusion. About a third of responses were of this sort. The second prediction is that if people overlook one or more

possibilities, then their conclusions should correspond to only *some* of the possibilities compatible with the premises. In fact, nearly all of the participants' erroneous conclusions were of this sort. Indeed, the most frequent errors were conclusions based on just a single possibility compatible with the premises. These errors cannot be attributed to blind guessing, because of the improbability of guessing so many conclusions compatible with the premises. People prefer to reason on the basis of a single model. Their erroneous conclusions are so hard to explain if they are relying on formal rules that no-one has so far devised such an explanation (pace Rips 1994, Braine & O'Brien 1998).

A simple way in which to prevent reasoners from being swamped by possibilities is to give them an extra premise that establishes the definite whereabouts of one of the persons, e.g.:

June is in England.
June is in Wales or Charles is in Scotland, but not both.
Charles is in Scotland or Kate is in Ireland, but not both.

What should then happen is that the interpretation of the first two premises yields only a single possibility:

June is in England Charles is in Scotland

The combination of this possibility with those for the third premise yields:

June is in England Charles is in Scotland Kate is *not* in Ireland

In this way, the number of possibilities that have to be kept in mind at any one time is reduced to one. The experiment included some problems of this sort, and they were easy. Diagrams can also improve performance with disjunctive problems, but not just any diagrams. They need to make the task of envisaging alternative possibilities easier (see Bauer & Johnson-Laird 1993).

Your working memory has a limited ability to hold models in mind. A superhuman intelligence, however, wouldn't be limited in this way. Its working memory would not be a bottleneck, and so it could reason with much more complex premises than you can. You don't realize your limitations because your social world is no more complicated than your ability to think about it—it couldn't be—and your reasoning about the physical world is good enough for you to survive.

4. The principle of parsimony

The model theory postulates that mental models are parsimonious. They represent what is possible, but not what is impossible, according to assertions. This principle of parsimony minimizes the load on working memory, and so it applies unless something exceptional occurs to overrule it. It

was introduced in Johnson-Laird & Savary (1999), who referred to it as the principle of "truth." This name is slightly misleading, and so I have changed it here. Some critics have thought that the principle means that mental models represent only those clauses mentioned in the premises. Such a view, however, would imply wrongly that sentences have the same models regardless of the connectives that occur in them.

The principle of parsimony is subtle because it applies at two levels. At the first level, mental models represent only what is possible. Consider, for example, how they represent the exclusive disjunction:

There is a circle or else there is a triangle but not both.

Its mental models represent the two possibilities:

○

△

where "○" denotes a model of the circle, "△" denotes a model of the triangle, and each horizontal line denotes a model of a separate possibility. Hence, the first row in this diagram represents the possibility described in the first clause in the sentence, and the second row represents the possibility described in the second clause. You will notice that two models of possibilities are more parsimonious than the four rows of a truth table, which represent both what is possible and what is impossible according to the premises.

The second level at which the principle of parsimony applies concerns individual models of possibilities: A mental model of a possibility represents a clause in the premises, whether it is affirmative or negative, only when the clause holds in that possibility. This principle is exemplified in the mental models of the disjunction above. The first model represents the possibility of a circle, but not the concurrent impossibility of a triangle. It contains no explicit information about the triangle. Likewise, the second model represents the possibility of a triangle, but not the concurrent impossibility of a circle. It contains no explicit information about the circle.

If you ask people to list what is possible given the preceding exclusive disjunction, they do indeed list a circle as one possibility, and a triangle as another possibility, and they say nothing about the status of the triangle in the first case or the status of the circle in the second case (Johnson-Laird & Savary 1999). Yet, they have not entirely forgotten what is impossible in a possibility that they represent. It is as though they made a mental footnote about it. But, the footnote is soon forgotten if they have to carry out a taxing piece of reasoning or if sentences contain several connectives.

Let's consider a different sentential connective, the conditional, which joins together two clauses using "if" and "then." Consider the assertion:

If there is a circle then there is a triangle.

You might ask: "And if there isn't circle, what then?" The answer is that there may or may not be a triangle. The conditional in its logical sense

is therefore compatible with three possibilities, which as usual I show on separate lines:

○ △
¬○ △
¬○ ¬△

where "¬" denotes negation. From adolescence or earlier, children list these possibilities, as do adults, when they are asked what is possible given a conditional (see, e.g. Barrouillet & Leças 1999, Barrouillet et al. 2000). However, because it's difficult to hold them all in mind, when individuals reason from a conditional, they focus on the possibility in which both the "if" clause, the *antecedent*, and the "then" clause, the *consequent*, occur. And so they construct the mental model:

○ △

But, if they were to construct only this model, then they would have represented a conjunction: There is a circle *and* there is a triangle. They realize that the antecedent needn't occur: There needn't be a circle. But, they defer the construction of an explicit model of this possibility. They construct only a model that has no explicit content. It acts as a "place holder" to remind them that there are other possibilities. The mental models of the conditional are accordingly:

○ △
...

where the ellipsis denotes the implicit model. Individuals should make a mental footnote that the possibilities represented in the implicit model are those in which the antecedent doesn't occur, i.e., there isn't a circle. If they retain this footnote, then they can flesh out their mental models into *fully explicit* models of the three possibilities. Now, you can understand why there is a difference in difficulty between the two conditional inferences with which I began the chapter. The easy inference follows at once from the mental models of the conditional, whereas the difficult inference does not. One way to make the difficult inference is to flesh out the mental models into fully explicit models; another way, which I will describe presently, is to make a supposition.

Just as there are two sorts of logical disjunction, inclusive and exclusive, so there are two sorts of logical conditional. You may have understood the conditional above to mean that if, *and only if*, there's a circle then there's a triangle. This interpretation is known as a biconditional, because it is equivalent to the assertion of two conditionals:

> If there is a circle then there is a triangle, and if there isn't a circle then there isn't a triangle.

The biconditional is compatible with only two possibilities:

○ △
¬○ ¬△

But, it has the same mental models as the regular conditional, except that the footnote states that the implicit model represents the possibility in which both clauses fail to hold. If you retain the footnote, then you should be able to flesh out your mental models into fully explicit models of the two possibilities. One reason that you will try to do so is if you are unable to draw a conclusion from your mental models.

Table 2 summarizes the mental models and the fully explicit models of sentences based on the *logical* meanings of the five principal sentential connectives. The ellipses represent implicit models, which serve as place holders representing other possibilities that as yet have no explicit content and that are constrained by mental footnotes. The fully explicit models flesh out mental models to represent all the clauses in the premises in all the possibilities.

5. Truth tables versus models

You should now understand the difference between truth tables and models. Truth tables represent truth values. Models represent possibilities. For example, the conjunction:

There is *not* a circle and there is a triangle

is represented by a truth table with four rows, which represents whether the atomic propositions are true or false. The only row for the conjunction that is true states in effect:

It is false that there is a circle and it is true that there is a triangle.

In contrast, the conjunction has a single mental model of a possibility:

¬ ○ △

Truth values are not possibilities, and the distinction matters in logic. When individuals refer to what is "true" or "false," or mentally represent these terms, they are at risk of paradox, as in the famous example from twentieth century logic:

This sentence is false.

If this sentence is true then it is false; if it is false then it is true. Of course, the sentence seems silly because it has no topic other than itself. Yet, logicians go to any lengths to avoid such paradoxes, because they are a symptom of an inconsistent system (see, e.g. Barwise & Etchemendy 1987). No risk of paradox occurs in referring to possibilities, e.g.:

This sentence is impossible.

The sentence merely makes a false claim about the grammar of English: "true" and "false" refer to the truth values of sentences, but "impossible" does not.

Table 2
The mental models and the fully explicit models for five sentential connectives

Connectives	Mental models		Fully Explicit models	
Conjunction:				
A and B:	A	B	A	B
Exclusive disjunction:				
A or B but not both:	A		A	¬ B
		B	¬ A	B
Inclusive disjunction:				
A or B or both:	A		A	¬ B
		B	¬ A	B
	A	B	A	B
Conditional:				
If A then B:	A	B	A	B
		...	¬ A	B
			¬ A	¬ B
Biconditional:				
If and only if A then B:	A	B	A	B
		...	¬ A	¬ B

Key: "¬" symbolizes negation, and "..." a wholly implicit model.

The difference between truth values and possibilities matters in psychology, because individuals respond differently to questions about truth and falsity than to questions about possibility and impossibility. For example, they tend to think that conditionals are *true* only in the case that both their clauses are true, but they are happy to list as *possible* all three cases in Table 2, corresponding to fully explicit models. Judgments of truth and falsity call for relating mental models to external possibilities in order to derive truth values. When individuals list possibilities, however, they have only to understand a sentence, and so they can flesh out their mental models into the three fully explicit models of a conditional.

6. Mechanisms of model building

The model theory postulates that humans have a natural disposition to think of possibilities. Alternative possibilities are represented as disjunctions of possibilities; and each model of a possibility represents a conjunction of affirmative and negative propositions. The theory as it applies to logical connectives therefore takes negation, conjunction, and inclusive disjunction, as fundamental. In this second part of the chapter, I am going to describe the mechanisms that construct models. These mechanisms have all been implemented in a computer program, and the program yields a surprising consequence, which I'll get to by and by. But, I begin with negation, and then proceed to connectives.

Here is a problem that turns out to be harder than it seems at first sight (see Barres & Johnson-Laird 2003). List the possibilities given the following assertion: It is not the case both that there is a circle and that there is a triangle. Why isn't the task trivial? The answer is that you don't know the answer, and so you have to infer it. You first have to work out what the unnegated sentence means:

There is a circle and there is a triangle.

It allows just one possibility:

○ △

The negative sentence rules out this possibility to leave its complement, which is all the other possible models based on the same two atoms and their negations. The first one that you're likely to think of is the mirror image of the preceding possibility:

¬ ○ ¬ △

Some individuals go no further, but you will realize that there are two other possibilities, in which one or other of the two shapes is missing:

¬ ○ △

○ ¬ △

In general, the way to infer the correct interpretation of a negative sentence is to take its atoms, and to work out all the possible combinations of them and their negations. You remove from these combinations those that are compatible with the unnegated sentence, and what remains is the answer: the possibilities compatible with the negative sentence. No wonder that people do not cope with the negation of compound sentences well. They tend to be better at negating a disjunction than a conjunction, perhaps because the former yields fewer models than the latter.

Individuals represent a set of alternative possibilities as a list of alternative models. Such a list corresponds to an inclusive disjunction. To combine two such sets of models according to any logical relation between them, calls only for negation, which I've described, and logical conjunction, which I'm

about to describe. When individuals interpret a set of premises, however, they construct a model of an initial clause or premise, and then update this model from the remaining information in the premises.

Let's consider a pair of premises that illustrate the main principles of conjunction:

If there is a triangle then there is a diamond.

There is a circle or else there is a triangle but not both.

Before I tell you what the resulting models are, you might like to think for yourself what possibilities are compatible with the two premises. Most people think that there are two: a triangle and a diamond, or a circle. The mental models of the first premise are:

△ ◊

...

The core of the interpretative process is to update these models by forming a conjunction of them with the models of the second premise. One possibility according to the second premise is that there is a circle, and so the system conjoins:

△ ◊ and ○

The triangle in the first model here occurs elsewhere in the models containing the circle, and so the interpretative system takes the absence of the triangle from the model containing the circle to mean that there is not a triangle. In effect, the conjunction becomes:

△ ◊ and ○ ¬ △

Because there is now a contradiction—one model contains a triangle and the other its negation—the result is a special null model (akin to the empty set), which represents propositions that are contradictory. It represents what is impossible. The conjunction therefore yields the null model:

nil

The system now conjoins the pair:

△ ◊ and △

The diamond doesn't occur elsewhere in the set of models containing the model of the triangle alone, and so the two models are compatible with one another. Their conjunction yields:

△ ◊

Similarly, the conjunction:

... and ○ yields ○

because the circle doesn't occur in the models containing the implicit model. The final conjunction:

... and △ yields nil

because the triangle does occur elsewhere in the models containing the implicit model, and so its absence in the implicit model is treated as akin to its negation. The mental models of the conjunction of the premises are accordingly:

Table 3
The mechanisms for conjoining pairs of mental models and pairs of fully explicit models

1. If one model contains a representation of a proposition, A, which is not represented in the other model, then consider the set of models of which this other model is a member. If A occurs in at least one of these models, then its absence in the current model is treated as its negation (go to mechanism 2); otherwise its absence is treated as its affirmation (go to mechanism 3). This mechanism applies only to mental models.
2. The conjunction of a pair of models containing respectively a proposition and its negation yield the null model, e.g.:
   A B and ¬A B yield nil.
3. The conjunction of a pair of models that are not contradictory yields a model containing all the elements of both models, e.g.:
   A B and B C yield A B C.
4. The conjunction of a null model with any model yields the null model, e.g.:
   A B and nil yield nil.

△ ◇
○

I have not shown the null models, because they do not represent possibilities. The two models of possibilities yield the valid conclusion:

There is a triangle and a diamond, or else there is a circle.

Table 3 summarizes the mechanisms for forming conjunctions of pairs of models.

The same mechanisms apply to the conjunction of fully explicit models. Here are the previous premises again:

If there is a triangle then there is a diamond.

There is a circle or else there is a triangle but not both.

Their mental models can be fleshed out to be fully explicit by a mechanism that uses mental footnotes, but I'll spare you the details. The fully explicit models of the conditional (see Table 1) are:

△ ◇
¬△ ◇
¬△ ¬◇

Because the disjunction has two models, there are six pair-wise conjunctions, but three of them are contradictions yielding the null model. The remaining pairs yield the following results:

△ ◇ ¬○
¬△ ◇ ○
¬△ ¬◇ ○

The same conclusion follows as before:

There is a triangle and a diamond or else there is a circle.

But, reasoners who rely on mental models will fail to think about the second of these possibilities. They should think that it is impossible to have the diamond and the circle. This prediction is typical of the model theory.

You can make suppositions when you reason, i.e., assumptions for the sake of argument (see, e.g. Byrne et al. 1995). Given a disjunction, such as:

There is a triangle on the board or there is a circle, or both.

you can make the supposition that there isn't a triangle on the board, and then infer as a consequence that in that case there is a circle on the board. You hold in mind a possibility, which in this case corresponds to the negation of an atom in the premise, and then treat it as though it was asserted categorically. You can then use the inferential mechanisms that I have already described. If you are prudent, you remember that any conclusion depends on a supposition, and take this fact into account in formulating a final conclusion. If a supposition leads to a contradiction (the null model), some individuals appreciate that the supposition is impossible granted the truth of the premises. The procedure is identical to the one that occurs in the construction of models of the following sort of conditional:

If A then both B and not B.

The conjunction, *B and not B*, yields the null model. The interpretation of the conditional calls for the conjunction of A and *nil*, which yields *nil* (see Table 3). What happens then depends on whether individuals are relying on mental models or fully explicit models. With mental models, there remains only the implicit model, which yields no conclusion. But, the fully explicit models of the conditional are:

A	nil
¬ A	nil
¬ A	¬ nil

The negation of *nil* in the third model yields the disjunction of the atoms that led to its construction, and so this conjunction yields the conclusion:

not A.

The corresponding principle in logic is known as reductio ad absurdum. In the model theory, it is a consequence of a mechanism that makes suppositions, and of reasoning from fully explicit models.

In a review of theories of conditionals, Evans & Over (2004) claimed that the model theory makes no use of suppositions, despite our several papers to the contrary (e.g. Byrne et al. 1995). They also argued that the model theory is truth functional, despite the arguments that I have summarized above. Their review is otherwise valuable. It is a pity that they mangle the model theory so badly, because it makes sense of phenomena that are otherwise puzzling for them, e.g., the difference that I described earlier between judgments of truth value and the listing of possibilities.

7. Superhuman reasoning

A computer program that I wrote to simulate the model theory can make inferences that are far beyond the ability of human reasoners working without benefit of logic. Only a superhuman intelligence, such as Hercule Poirot (Agatha Christie's famous fictional detective), could solve the following problem without paper and pencil:

> Who helped to murder Mr. Ratchett on the Orient Express?
> If Pierre helped if Dr. Constantine did then Greta helped too.
> If not both Harriet and Hector helped then Dr. Constantine didn't help.
> Greta didn't help or Dr. Constantine did.
> Harriet didn't help or the Princess Drago-Miroff, Mary, and Colonel Arbuthnot all helped.
> If Hector or Mary helped then Pierre helped or Colonel Arbuthnot didn't help.
> So, who helped, who didn't help, and for whom is it impossible to say?

There are eight atomic propositions in the premises, and so their truth table has 256 rows. Likewise, there are multiple models, but if you build them up premise by premise, the final result is a single model. It shows that all eight individuals helped to commit the murder. In many other inferences, of course, the premises yield multiple models, but an algorithm exists for drawing parsimonious conclusions that describe them (see Johnson-Laird & Byrne 1991, Ch. 9).

8. Some illustrative inferences

To illustrate the model theory and its predictions, I am going to consider some inferences that human reasoners *can* make. The first inference is:

> Either Jane is kneeling by the fire and she is looking at the TV or else Mark is standing at the window and he is peering into the garden.
> Jane is kneeling by the fire.
> Does it follow that she is looking at the TV?

Most people say: "yes" (Walsh & Johnson-Laird 2004). A second inference has the same initial premise, but it is followed instead by the categorical denial:

> Jane is not kneeling by the fire.

and the question is:

> Does it follow that Mark is standing at the window?

Again, most individuals say: "yes". Let's see what the theory predicts.

The first premise in both inferences is the same exclusive disjunction of two conjunctions. The theory predicts that individuals should rely on mental models. Hence, they should interpret the first conjunction, Jane is kneeling by the fire and she is looking at the TV, and build a model representing this possibility, which I will abbreviate as follows:

Jane: kneeling looking

They should build an analogous model of the second conjunction:

Mark: standing peering

These two models must now be combined according to an exclusive disjunction. An exclusive disjunction has two mental models, which represent the two conjunctions only in the possibilities in which they hold:

Jane: kneeling looking

Mark: standing peering

For the first inference, the conjunction of the categorical premise:

Jane is kneeling

with the first model of the disjunction yields:

Jane: kneeling looking

Its conjunction with the second model of the disjunction yields the null model. Hence, the premises yield only the model:

Jane: kneeling looking

and so individuals should respond: yes, Jane is looking at the TV. This analysis may strike you as obvious.

In fact, the inference is a fallacy. The principle of parsimony postulates that individuals normally represent what is possible, but not what is impossible. When I first wrote the computer program to simulate the theory, and inspected its output for a certain problem, I thought that there was a bug in the program. I searched for the bug for half a day, before I realized that the program was correct, and the error was in my thinking. What the program revealed is the discrepancy between mental models and fully explicit models. The theory therefore predicted that individuals should reason in a fallacious way for certain inferences. Indeed, the fallacies turn out to be so compelling in some cases that they resemble cognitive illusions, and so my colleagues and I refer to them as "illusory" inferences.

If you succumbed to the illusion, then you are in the company of Clare Walsh and myself. We studied these inferences, but it took us a couple of days to realize that they were illusory, and that was *after* the discovery of other sorts of illusions. The fully explicit models of the exclusive disjunction reveal the correct conclusion:

Jane:	kneeling	looking	Mark:	¬ standing	¬ peering
Jane:	kneeling	looking	Mark:	¬ standing	peering
Jane:	kneeling	looking	Mark:	standing	¬ peering
Jane:	¬ kneeling	¬ looking	Mark:	standing	peering

Jane: ¬ kneeling looking Mark: standing peering
Jane: kneeling ¬ looking Mark: standing peering

When one conjunction is true, the other conjunction is false, and you will remember from my earlier account that there are three ways in which a conjunction can be false. The categorical premise that Jane is kneeling rules out the fourth and fifth possibilities. But, contrary to the illusory inference, it leaves one possibility—the sixth one—in which Jane is kneeling but not looking at the TV. That is why the illusory inference is invalid. Granted that Jane is kneeling, it does not follow that she is looking at the TV.

The second problem that I described has the categorical premise that Jane is not kneeling by the fire, and poses the question of whether it follows that Mark is standing by the window. Most people respond, "yes", which is a conclusion supported by the mental models shown above. The fully explicit models show that this inference *is* valid. The categorical premise eliminates all but the fourth and fifth models, and in both of them Mark is standing by the window. Our main experiment examined a series of illusory inferences and control problems of this sort. The participants were much more likely to respond correctly to the control problems (78% correct) than to the illusory problems (10% correct): 34 of the 35 participants showed this difference.

Illusory inferences occur in many domains, including reasoning with quantifiers (Yang & Johnson-Laird 2000*a*,*b*), deontic reasoning (Bucciarelli & Johnson-Laird 2005), and assessing whether or not sets of assertions are consistent (Johnson-Laird et al. 2004). I will describe two more examples.

The first example (from Goldvarg & Johnson-Laird 2000) calls for reasoning about what is possible:

> Only one of the following premises is true about a particular hand of cards:
>
> There is a king in the hand or there is an ace, or both.
> There is a queen in the hand or there is an ace, or both.
> There is a jack in the hand or there is a 10, or both.
>
> Is it possible that there is an ace in the hand?

The model theory postulates that individuals consider the possibilities for each of the three premises. That is, they assume that the first premise is the one that is true, and consider the consequences; then they assume that the second premise is the one that is true and consider the consequences, and then they assume that the third premise is the one that is true and consider the consequences. However, because the question asks only whether an ace is *possible*, they can stop as soon as they find a premise that allows the presence of the ace in the hand. What is wrong with this procedure? The answer is that when individuals consider the truth of one premise, they should also consider the concurrent *falsity* of the other two premises. But, that is exactly what the principle of parsimony predicts they will not do.

For the first premise, they accordingly consider three models, which each correspond to a possibility given the truth of the premise:

king

ace

king ace

Two of the models show that an ace is possible. Hence, on the basis of this premise alone individuals should respond, "yes." The second premise supports the same conclusion. The third premise is compatible with it. In fact, it is an *illusion of possibility*: reasoners infer wrongly that a card is possible. If there were an ace, then two of the premises would be true, contrary to the rubric that only one of them is true. The same strategy, however, yields a correct response to a control problem in which only one premise refers to an ace. A problem to which reasoners should respond "no," and thereby succumb to an *illusion of impossibility*, can be created by replacing the two occurrences of "there is an ace" in the premises above with, "there is not an ace." Its control problem contains only one premise with the clause, "there is not an ace."

Figure 1 presents the results of an experiment in which we gave students 16 inferences, four of each of the four sorts. Half of the illusions were based on disjunctive premises, and half were based on conditionals. The participants' confidence in their conclusions did not differ reliably from one sort of problem to another. As the Figure shows, they were very susceptible to the illusions but performed well with the control problems, and the illusions of possibility were more telling than those of impossibility. To infer that a situation is impossible calls for a check of every model, whereas to infer that a situation is possible does not, and so reasoners are less likely to make the inference of impossibility. This difference also occurs in problems that are not illusory (Bell & Johnson-Laird 1998).

With hindsight, it is surprising that nearly everyone responded "yes" to the first of the problems above, because it seems obvious that an ace renders two of the premises true. We therefore carried out a replication with two groups of participants, and half way through the experiment, we told one group to check whether their conclusions met the constraint that only one of the premises was true. This procedure had the advantage that the participants did not have to envisage the circumstances in which the premises did not hold. The group that received the special instruction was thereafter much less likely to commit the fallacies (Goldvarg & Johnson-Laird 2000).

If the preceding illusions result from a failure to reason about what is false, then any manipulation that emphasizes falsity should reduce them. The rubric, "Only one of the following two premises is *false*," did reduce their occurrence (Tabossi et al. 1998), as did the prior production of false instances of the premises (Newsome & Johnson-Laird 1996).

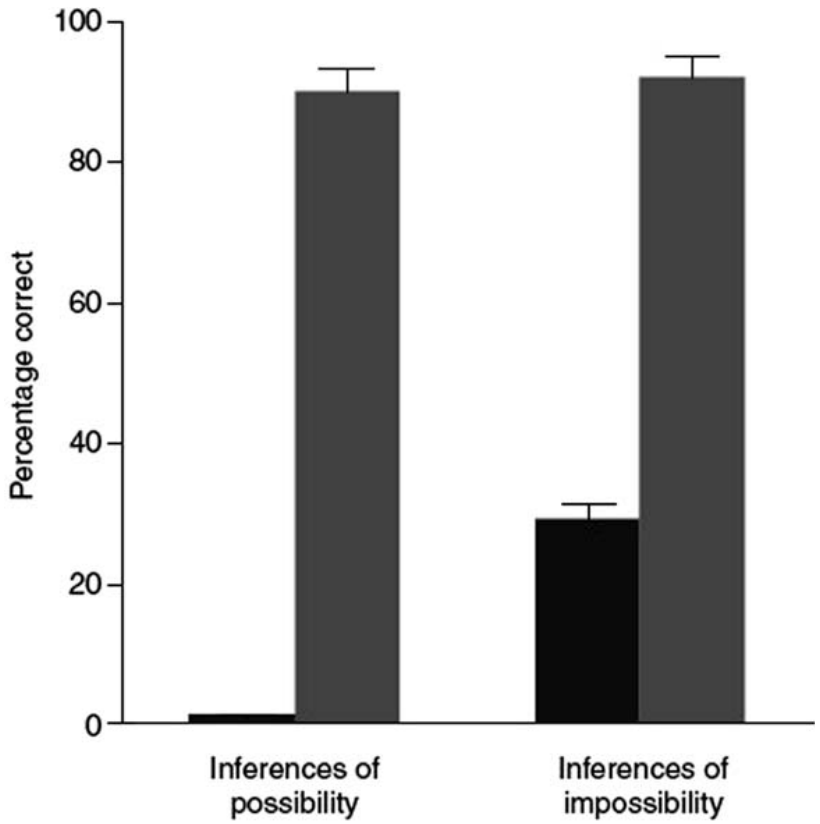


Fig. 1. The percentages of correct responses to fallacious inferences that are illusory and their control problems (based on Goldvarg and Johnson-Laird, 2000)

The second example of an illusion is very compelling. The rubric, "one of these assertions is true and one of them is false," is equivalent to an exclusive disjunction between two assertions. Consider this problem, which is based on an exclusive disjunction:

> Suppose you know the following about a particular hand of cards:
> If there is a jack in the hand then there is a king in the hand, or else if there isn't a jack in the hand then there is a king in the hand.
> There is a jack in the hand.
> What, if anything, follows?

Nearly everyone—experts and novices alike—infers that there is a king in the hand (Johnson-Laird & Savary 1999). Yet, it is a fallacy granted a disjunction, exclusive or inclusive, between the two conditional assertions. The disjunction entails that one or other of the two conditionals could be

false; and if one of them is false, then there may not be a king in the hand. Suppose, for instance, that the first conditional is false. There could then be a jack but *not* a king—a judgment with which most individuals concur (see, e.g. Oaksford & Stenning 1992). And so the inference that there is a king is invalid: the conclusion could be false.

An experiment examined the preceding problem and another illusion, and compared them with two control problems in which the neglect of false cases should not impair performance (Johnson-Laird & Savary 1999). The participants committed both fallacies in 100 percent of cases, and yet drew valid inferences for the control problems in 94 percent of cases. The participants were again confident in both their illusory conclusions and their correct control conclusions.

Because so many expert psychologists have succumbed to illusory inferences, we have accumulated many putative explanations for them. For example, the premises may be so complex, ambiguous, or odd, that they confuse people, who, as a result, commit a fallacy. This hypothesis overlooks the fact that the participants are very confident in their conclusions, and that the control problems are equally complex. Likewise, when the illusions and controls are based on the *same* premises, but different questions in the form of conjunctions, participants still commit the fallacies and get the control problems correct (Goldvarg & Johnson-Laird 2000).

Other putative explanations concern the interpretation of conditionals. Individuals make the illusory inference with problems of this sort:

One of the following assertions is true and one of them is false:

If there is a jack then there is a king.

If there isn't a jack then there is a king.

This assertion is definitely true:

There is a jack.

Naïve individuals understand that a conditional, such as:

If there is jack then there is a king.

is false in the case that there is jack but not a king. They also understand that the rubric to this problems mean that one conditional is true and the other conditional is false. Hence, on their own account they should refrain from inferring that there is a king. The analysis depends on nothing else. However, even if some special factors exacerbate the illusions with conditionals, other illusions occur with problems that do not contain conditionals, such as the problem with which I started this section of the chapter.

Many other robust phenomena in reasoning appear to arise from the principle of parsimony and the resulting neglect of what is impossible or false. They include the results of Wason's "selection" task in which individuals fail to grasp the relevance of an instance of a false consequent to testing the truth or falsity of a conditional (see, e.g. Wason 1966, Wason & Johnson-Laird 1972).

9. Conclusions

This chapter has explained the mechanisms that construct models based on the logical interpretation of connectives. These models do not represent truth values, but sets of possibilities. Individuals adopt a variety of strategies to cope with reasoning problems, e.g., they may be guided by a given conclusion, they may work forwards from the premises, they may make a supposition, and so on (Van der Henst et al. 2002, Johnson-Laird & Hasson 2003). But, regardless of strategy, inferences that depend on a single model are easier than those that depend on multiple models.

Mental models abide by the principle of parsimony: They represent only possibilities compatible with the premises, and they represent clauses in the premises only when they hold in a possibility. Fully explicit models represent clauses when they do not hold too. The advantage of mental models over fully explicit models is that they contain less information, and so they are easier to work with. But they can lead reasoners astray. The occurrence of these systematic and compelling fallacies is shocking. The model theory predicts them, and they are a "litmus" test for mental models, because no other current theory predicts them. They have so far resisted explanation by theories of reasoning based on formal rules of inference, because these theories rely on valid rules. For several years, my former colleague Yingrui Yang has sought an explanation based on a revised formal rule theory, but he has yet to succeed. To reason only about what is possible is a sensible way to cope with limited processing capacity, but it does lead to illusions. Yet, it does not imply that people are irredeemably irrational. The fallacies can be alleviated with preventative methods. Otherwise, however, reasoners remain open to the illusion that they grasp what is in fact beyond them.

References

Barres, P. & Johnson-Laird, P. (2003), 'On imagining what is true (and what is false)', *Thinking & Reasoning* **9**, 1–42.

Barrouillet, P., Grosset, N. & Leças, J. F. (2000), 'Conditional reasoning by mental models: chronometric and developmental evidence', *Cognition* **75**, 237–266.

Barrouillet, P. & Leças, J.-F. (1999), 'Mental models in conditional reasoning and working memory', *Thinking & Reasoning* **5**, 289–302.

Barwise, J. (1993), 'Everyday reasoning and logical inference', *Behavioral and Brain Sciences* **16**, 337–338.

Barwise, J. & Etchemendy, J. (1987), *The Liar: An Essay in Truth and Circularity*, Oxford University Press, New York.

Bauer, M. & Johnson-Laird, P. (1993), 'How diagrams can improve reasoning', *Psychological Science* **4**, 372–378.

Bell, V. & Johnson-Laird, P. (1998), 'A model theory of modal reasoning', *Cognitive Science* **22**, 25–51.

Berry, G. (1952), Peirce's contributions to the logic of statements and quantifiers, *in* P. Wiener & F. Young, eds, 'Studies in the Philosophy of Charles S. Peirce', Harvard University Press, Cambridge, MA.

Boolos, G. & Jeffrey, R. (1989), *Computability and Logic*, 3rd edn, Cambridge University Press, Cambridge.

Braine, M. & O'Brien, D., eds (1998), *Mental Logic*, Erlbaum, Mahwah, NJ.

Bucciarelli, M. & Johnson-Laird, P. (2005), 'Naïve deontics: a theory of meaning, representation, and reasoning', *Cognitive Psychology* **50**, 159–193.

Byrne, R. (2005), *The Rational Imagination: How People Create Alternative to Reality*, MIT Press, Cambridge, MA.

Byrne, R., Handley, S. & Johnson-Laird, P. (1995), 'Reasoning from suppositions', *Quarterly Journal of Experimental Psychology* **48A**, 915–944.

Evans, J., Newstead, S. & Byrne, R. (1993), *Human Reasoning: The Psychology of Deduction*, Erlbaum, Hillsdale, NJ.

Evans, J. & Over, D. (2004), *If*, Oxford University Press, Oxford.

Garnham, A. (2001), *Mental Models and the Representation of Anaphora*, Psychology Press, Hove, East Sussex.

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

Goldvarg, Y. & Johnson-Laird, P. (2000), 'Illusions in modal reasoning', *Memory & Cognition* **28**, 282–294.

Jeffrey, R. (1981), *Formal Logic: Its Scope and Limits*, 2nd edn, McGraw-Hill, New York.

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

Johnson-Laird, P. & Byrne, R. (1991), *Deduction*, Erlbaum, Hillsdale, NJ.

Johnson-Laird, P. & Byrne, R. (2002), 'Conditionals: A theory of meaning, pragmatics, and inference', *Psychological Review* **109**, 646–678.

Johnson-Laird, P., Byrne, R. & Schaeken, W. (1992), 'Propositional reasoning by model', *Psychological Review* **99**, 418–439.

Johnson-Laird, P., Girotto, V. & Legrenzi, P. (2004), 'Reasoning from inconsistency to consistency', *Psychological Review* **111**, 640–661.

Johnson-Laird, P. & Hasson, U. (2003), 'Counterexamples in sentential reasoning', *Memory & Cognition* **31**, 1105–1113.

Johnson-Laird, P., Legrenzi, P., Girotto, V., Legrenzi, M. & Caverni, J.-P. (1999), 'Naive probability: a mental model theory of extensional reasoning', *Psychological Review* **106**, 62–88.

Johnson-Laird, P. & Savary, F. (1999), 'Illusory inferences: A novel class of erroneous deductions', *Cognition* **71**, 191–229.

Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman, San Francisco.

Medvedev, Z. A. (1990), *The Legacy of Chernobyl*, W.W. Norton, New York.

Metzler, J. & Shepard, R. (1982), Transformational studies of the internal representations of three-dimensional objects, *in* R. Shepard & L. Cooper, eds, 'Mental Images and Their Transformations', MIT Press, Cambridge, MA, pp. 25–71.

Newsome, M. & Johnson-Laird, P. (1996), An antidote to illusory inferences, *in* 'Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society', Erlbaum, Mahwah, NJ, p. 820.

Oakhill, J. & Garnham, A., eds (1996), *Mental Models in Cognitive Science*, Psychology Press, Hove, Sussex.

Oaksford, M. & Stenning, K. (1992), 'Reasoning with conditionals containing negated constituents', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **18**, 835–854.

Polk, T. & Newell, A. (1995), 'Deduction as verbal reasoning', *Psychological Review* **102**, 533–566.

Rips, L. (1994), *The Psychology of Proof*, MIT Press, Cambridge, MA.

Stevenson, R. (1993), *Language, Thought and Representation*, Wiley, New York.

Tabossi, P., Bell, V. & Johnson-Laird, P. (1998), Mental models in deductive, modal, and probabilistic reasoning, *in* C. Habel & G. Rickheit, eds, 'Mental Models in Discourse Processing and Reasoning', John Benjamins, Berlin.

Van der Henst, J.-B., Yang, Y. & Johnson-Laird, P. (2002), 'Strategies in sentential reasoning', *Cognitive Science* **26**, 425–468.

Walsh, C. & Johnson-Laird, P. (2004), 'Co-reference and reasoning', *Memory & Cognition* **32**, 96–106.

Wason, P. (1966), Reasoning, *in* B. Foss, ed., 'New Horizons in Psychology', Penguin, Harmondsworth, Middx.

Wason, P. & Johnson-Laird, P. (1972), *The Psychology of Deduction: Structure and Content*, Harvard University Press/Batsford, Cambridge, MA/London.

Wittgenstein, L. (1922), *Tractatus Logico-Philosophicus*, Routledge & Kegan Paul, London.

Yang, Y. & Johnson-Laird, P. (2000*a*), 'How to eliminate illusions in quantified reasoning', *Memory & Cognition* **28**, 1050–1059.

Yang, Y. & Johnson-Laird, P. (2000*b*), 'Illusory inferences with quantified assertions: How to make the impossible seem possible, and vice versa', *Memory & Cognition* **28**, 452–465.

This Page is Intentionally Left Blank

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Interaction of Knowledge and Working Memory in Reasoning About Relations

André Vandierendonck, Vicky Dierckx & Hannelore Van der Beken

Department of Experimental Psychology, Ghent University [1]

Abstract
This chapter defends the thesis that relational reasoning is based on an interaction of knowledge represented in semantic and episodic long-term memory on the one hand and temporary information maintained in working memory on the other hand. Three lines of evidence relevant to this interaction are considered. First, it is shown that reasoners seem to have a preference for a visuo-spatial representation of the temporary mental models when this representational format is advantageous because there is a need to represent structural information or because the relations are easily represented in this format. The second line of evidence shows that apart from believability of the premises and believability of the conclusions, also believability of the model(s) described by the premises plays a role. In a third part, transitive (order relations) and intransitive relational problems (genealogical relations) are compared in order to clarify the role of constraints on the relational representations on reasoning. The results obtained in these three lines of investigation support the view that (relational) reasoning is based on a tight interplay of knowledge representations in long-term memory and temporary models in working memory. Some discussion is devoted to the possibility that this interaction is mediated by Baddeley's concept of episodic buffer.

[1] E-mail: Andre.Vandierendonck@ugent.be

1. Introduction

The present chapter develops the thesis that deductive reasoning about relations involves an interplay of knowledge available in semantic and episodic long-term memory on the one hand and temporary information maintained in working memory on the other hand. First, we specify the nature of relational reasoning and show how the mental models theory of reasoning (Johnson-Laird 1983) applies to this kind of reasoning. We also address the question how working memory and its different components support relational reasoning. Next, attention will be focused on three specific ways in which knowledge may affect reasoning performance. A first theme addresses the visuo-spatial basis of relational reasoning. The second theme focuses on how activation of relevant background knowledge available in cognitive schemata such as scripts may help or impair reasoning about temporal relations. Finally, we discuss reasoning about family relations. These constitute a set of relations which have a rather rich semantic content imposing many constraints on the inferences that can be made.

1.1. Relational reasoning

From the information that `Adam was born before Bob` and that `Bob was born before Carol`, we can easily infer that `Adam was born before Carol`. In this example, an inference must be made about the relation between `Adam` and `Carol` and this is quite straightforward since "born before" implies a temporal order relation which is transitive. Most of the research on relational reasoning has used such transitive problems. It is not always the case, however, that the premises support a valid conclusion, as can be seen in the following example. From `Dedre was born before Eric` and `Dedre was born before Fay`, it is not possible to decide whether `Eric was born before Fay` or vice versa. These are indeterminate problems and are more difficult to solve. Research based on problems with three terms as in the examples given (three-term series), has shown that people solve such problems by integrating all the relations into a single spatial array either with the largest values at the top and the smallest at the bottom or in a left-to-right orientation with the largest values to the right (e.g. De Soto et al. 1965, Huttenlocher 1968, Potts 1974). Although an alternative view was proposed and defended by Clark (1969, 1971), the evidence collected over the years shows that the *spatial array view* yields the best summary and explanation of the findings (see e.g. chapter 6 in Evans et al. 1993). The relational content does not seem to matter much and the view applies equally well for determinate as for indeterminate problems based on transitive relations (Barclay 1973, Foos et al. 1976, Huttenlocher 1968,

Maybery et al. 1986, Mynatt & Smith 1977, Potts 1974, 1976, Sternberg 1980, 1981).

1.2. Mental models

A spatial array is nothing else than a mental representation in which the "objects" are given a place along a spatial line. It is a (mental) model representing the information given in the premises. In this sense, the spatial array view is a special case of the *mental models theory* developed by Johnson-Laird and colleagues (Goodwin & Johnson-Laird 2005, Johnson-Laird 1983, Johnson-Laird & Byrne 1989). The main thesis of this theory is that deductive reasoning is a *meaning-based*, rather than a rule-based, process. Basically, reasoning follows a three-stage course consisting of an interpretation of the premises, the formulation of a tentative conclusion and a search for counter-examples.

Stages in reasoning. The first stage yields an *interpretation of the information given in the premises.* In addition to a comprehension of each of the statements, this requires an understanding of the relationship expressed in each premise. The premise information is then integrated into one or more models, whereby each model is a representation of a possible situation in the world, given the information in the premises. With the premises **`Glenda is smarter than Howard`** and **`Glenda is smarter than Igor`**, an initial model could be **`Glenda-Howard-Igor`**.

In the next phase, the models are used to *generate a tentative conclusion.* From the model constructed in the last example, a tentative conclusion would be that **`Howard is smarter than Igor`**. Because the conclusion is only tentative, it is necessary to check whether the conclusion really holds. This is achieved in a third phase, by *searching for counter-examples.* In our example, the reasoner might wonder whether it is not possible that **`Igor is smarter than Howard`**. This corresponds to another model, namely **`Glenda-Igor-Howard`** which appears to be consistent with the premises also. However, this second model is not consistent with the tentative conclusion. Since there are now two models representing the premises and the tentative conclusion is consistent with only one of the models, this conclusion cannot be valid. Actually, the reasoner is now able to recognize that the premises allow to construct two models that are contradictions of each other. As a consequence, with the premises given in the last example, the reasoner can conclude that no valid conclusion can be obtained.

Implicit models. A basic tenet of the mental models theory is that in the first stage an initial model is constructed and that room is left for further elaboration of the representations by including an implicit model. Whilst this is an important and central thesis in the theory which has

found support in studies of syllogistic and conditional reasoning (see e.g., Johnson-Laird 1999, Johnson-Laird & Byrne 1989), in research on relational reasoning, on the contrary, the assumption that the reasoner immediately constructs all relevant models has been prevalent (e.g., Evans et al. 1993, Johnson-Laird & Byrne 1989). However, Vandierendonck et al. (2004) have shown that the latter assumption is probably incorrect. They used four-term series problems in which the premises could be consistent with one, two or three different models. The findings supported the hypothesis that reasoners construct an integrated representation of all the premise information and that in multi-model problems an *annotation* is made which is further fleshed out when the need arises. This elaboration of the annotation does not result in the construction of more than one model, however; instead, the reasoners seem to construct a forked array which has been labeled "isomeric model" (Schaeken et al. in press). For example, the premises **John is larger than Kate**, **Kate is larger than Lee**, and **John is larger than Meg**, are consistent with three different models, namely **John-Kate-Lee-Meg**, **John-Kate-Meg-Lee**, and **John-Meg-Kate-Lee**. Instead of constructing or elaborating all these models, the reasoners start with constructing an annotated model such as **John-(Meg)-Kate-Lee** and eventually work this out into a construction like $\texttt{John}-\frac{\texttt{Kate-Lee}}{\texttt{Meg}}$, which indicates that **Meg** may occur at any position after **John**.

Working memory. According to the mental models theory, a model is constructed to integrate information extracted from the premises as a possible representation of what is given to be true.[2] By definition, a model is a *temporary construction*, needed for the derivation of a tentative conclusion and for supporting the search for counterexamples. Once these processes come to a closure, the need for maintaining the model(s) no longer exists. The designated medium for the maintenance of temporary supporting information, such as models, is working memory. Indeed, another assumption of the mental models theory is that *models are constructed in working memory* and since working memory has a limited capacity, the more models that have to be maintained simultaneously in working memory, the more likely it is that information will be lost and that errors will be made and the more time the reasoner will need to arrive at a conclusion.

Within the different conceptualizations of working memory (see e.g., Miyake & Shah 1999, for a broad overview), the model of Baddeley & Hitch (1974) provides a useful framework for the study of the interaction of different cognitive tasks with working memory (see also, Baddeley 1986, 2000, Baddeley & Logie 1999). The strength of the model is that it

[2] This formulation may be perceived as strange. Nevertheless, it is related to a central tenet of the mental models theory, namely that the reasoners only construct models on the basis of information that is given as being true.

conceptualizes working memory as a multicomponential system in which a number of automatically operating processes are monitored by an executive controller. Originally, two subsidiary systems were postulated, one for the maintenance of phonologically coded materials, the *phonological loop*, and one for the maintenance of visuo-spatially coded materials, the *visuo-spatial sketch pad*. According to recent developments, both the phonological loop (Baddeley 1986) and the visuo-spatial sketch pad (Logie 1995) are considered to consist of a system for the passive maintenance of modality-specific information and a looping device for refreshing the information kept in the store. Recently, a third subsidiary system has been proposed, the *episodic buffer*, which has the task of integrating information from both other subsidiary components with the contents of episodic memory (Baddeley 2000). All these subsidiary systems are supervised by a *central executive* which is similar to the *supervisory attentional system* described by Norman & Shallice (1986).

Viewed within the context of this general framework, there is little doubt that reasoning calls on the executive controller, at least to the extent that reasoning involves the evaluation of tentative conclusions and the search for counterexamples (see Evans 2000, for a discussion of this issue). Indeed, several studies using a dual-task methodology have shown that reasoning is impaired when it is performed concurrently with a demanding secondary task (see e.g., Klauer et al. 1997, Meiser et al. 2001, Vandierendonck & De Vooght 1997). In a similar vein, there is evidence that the other working memory components also play a role in deductive reasoning in providing storage for the maintenance of the models. However, as will be explained later on in this chapter, it seems to depend on several factors such as the kind of reasoning task, the number of models to be constructed, the size of the models, etc. whether and to what extent reasoning calls on the phonological and the visuo-spatial subsystems (see e.g., Duyck et al. 2003, Gilhooly et al. 1993, 1999, Klauer et al. 1997, Meiser et al. 2001, Toms et al. 1993).

Long-term Memory. The construction of mental models, being a meaning-driven process, also heavily relies on *long-term memory*. Because reasoning is assumed to rely on a semantic analysis of the premises, reasoning is a process based also on accumulated knowledge. In fact, long-term memory support may intrude in at least four different steps of the reasoning process.

1. The premises must be comprehended. Apart from a process of language comprehension, this also involves recollection of contexts that are relevant to the meaning of the premise sentences. Available information may provide "additional premises" so that particular inferences are facilitated while others may be suppressed (e.g., Byrne 1989). In a similar vein, knowledge and beliefs may provide models or conclusions that are difficult to ignore and that do affect the rep-

resentations reasoning is based on (e.g., belief biases, Revlin et al. 1980).

2. In the construction of models, the information from several premises must be integrated. Again, available knowledge may play a role to succeed in this task. This is one of the issues that will be further developed in the present chapter.
3. General knowledge will support the process of conclusion generation. Knowledge provides possible conclusions. It has long been known, that in the absence of knowledge, as in reasoning problems with symbolic entities, reasoning is less biased. The present chapter also presents evidence that knowledge embedded in schemata or scripts may provide alternative interpretations of premises in relational reasoning.
4. In the search for counterexamples, the availability of relevant and useful information retrieved from long-term memory may again affect the outcome of the reasoning process. If a conclusion contradicts knowledge, an alternative is readily available and may contribute to the observation of belief biases (e.g., Byrne et al. 2000, Newstead et al. 1992).

1.3. Coupling of working memory and long-term memory

For a long time, the role of long-term memory in reasoning has been a nuisance to reasoning theorists, because the rule-based views on reasoning have always had difficulty to explain how the application of logical rules could be intruded by knowledge (e.g., Henle 1962). For a meaning-based theory, as the mental models theory is, these effects are part and parcel of reasoning. Interestingly, as reasoning is supposed to be mediated by working memory, such a view also implies that at some point working memory and long-term memory should interact on the way to obtain a solution to a reasoning problem. Even though, thus far, the mental models theory does not provide a computational view on deductive reasoning, this particular interaction enables a first approach to the specification of a processing account of model-based reasoning. Before this can be achieved, however, we need more knowledge about how this interaction affects reasoning. The present chapter is an attempt to bring together such essential information along three lines, namely the visuo-spatial nature of relational reasoning, the role of the believability of relational models and the specificity of relational information. Each of these three aspects is elaborated next.

2. Visuo-spatial basis

In the introduction, we reviewed already the evidence from relational reasoning studies showing that (a) from the start an integrated representation of the premise information is constructed, (b) that this representation may be left implicit by adding an annotation that can be further elaborated, and (c) that after elaboration a forked structure may result. This all very strongly suggests a spatial basis for the representation of such models in line with the original spatial array theory. If this observation is correct, one should expect that relational reasoning strongly relies on visuo-spatial working memory, especially when the terms cannot be represented as a single ordered series. Indeed, if only one-model problems are involved, the problems can be represented by a single order of tokens (terms, names, ...) and this can be achieved by the phonological loop because all what is needed is a memory for a string of tokens. This may be particularly easy when the terms can be represented by their first letter so that a pseudo-word can be used to represent the model or when the terms are all short words that fall easily within the storage capacity of the phonological loop (Dierckx et al. 2003).

There may be at least two reasons why with relational problems, reasoners would rely on the visuo-spatial sketch pad rather than the phonological loop for constructing and maintaining the mental model(s). A first reason is that the problem cannot be represented by a single string because some form of elementary structure is present in the representation. A forked representation such as $\texttt{John}-\frac{\texttt{Kate}-\texttt{Lee}}{\texttt{Meg}}$ thus would be more easy to maintain in a visuo-spatial code. A second reason is that some relations are more easily mapped on a spatial display than other ones. It is evident that spatial relations such as "left of," "above," "behind" are more easily represented in a spatial array than relations of other types. However, temporal relations, such as "before," "at the same time," are also easy to map on a spatial array, because we are used to this form of representation (time lines, clocks, etc.). For still other types of relations, the mapping may be more difficult, but still possible.

Whereas the first of these reasons for using the visuo-spatial sketch pad to construct models mainly capitalizes on the structure and the complexity of the models, the second reason is a more intrinsic one based on how easy the meaning of the relation is translated into a spatial representation. In what follows, we will review empirical evidence relevant for both aspects.

2.1. Visuo-spatial working memory in spatial and temporal reasoning

To address the first issue, we can refer to a study of Vandierendonck & De Vooght (1997). These authors investigated the involvement of three working memory components in four-term linear reasoning tasks, namely the phonological loop, the visuo-spatial sketch pad and the central executive. In the second experiment of their study, fourty-four participants were randomly assigned to four dual-task conditions, namely control (reasoning only), articulatory suppression (reasoning while continuously producing a fixed string of four digits at a fast rate), matrix tapping (reasoning while continuously tapping the four corners of the numeric keypad at a fast rate) and random interval repetition (reasoning while "shadowing" a series of randomly spaced time intervals by tapping a key[3]). In comparison to the participants in the single-task control condition, the participants in the articulatory suppression condition did not have the possibility to verbally rehearse the premises. If model construction uses the phonological loop, reasoning performance of these participants should be drastically impaired, because it would be very difficult to maintain the information in working memory. This inference is based on evidence showing that articulatory suppression interferes with short-term memorization of verbal information (see e.g., Baddeley et al. 1984). Compared to the participants in the control condition, those in the matrix tapping condition should show impaired performance if reasoning is based on a visuo-spatial representation of the premise information. Again previous research has shown that active and passive movements interfere with visuo-spatial working memory (see e.g., Quinn 1994, Smyth & Scholey 1992, 1994). Finally, in comparison to the participants in the control condition, those in the random interval repetition condition should be impaired if reasoning relies on executive control processes.

The premises were either presented at a fixed speeded rate (3 s per premise) or they were presented in a self-paced way with registration of the time taken to read and to process each premise. It was assumed that in the speeded condition, the participants would not have enough time to integrate all the information in the premises in a model and hence performance should be worse than in the self-paced reading condition. As expected, accuracy was poorer in the speeded reading condition (51% correct) than in the self-paced condition (61% correct), and reasoning performance was impaired in each of the three dual-task conditions.

[3] In fact, this is a continuous simple reaction-time task with randomly selected interstimulus intervals.

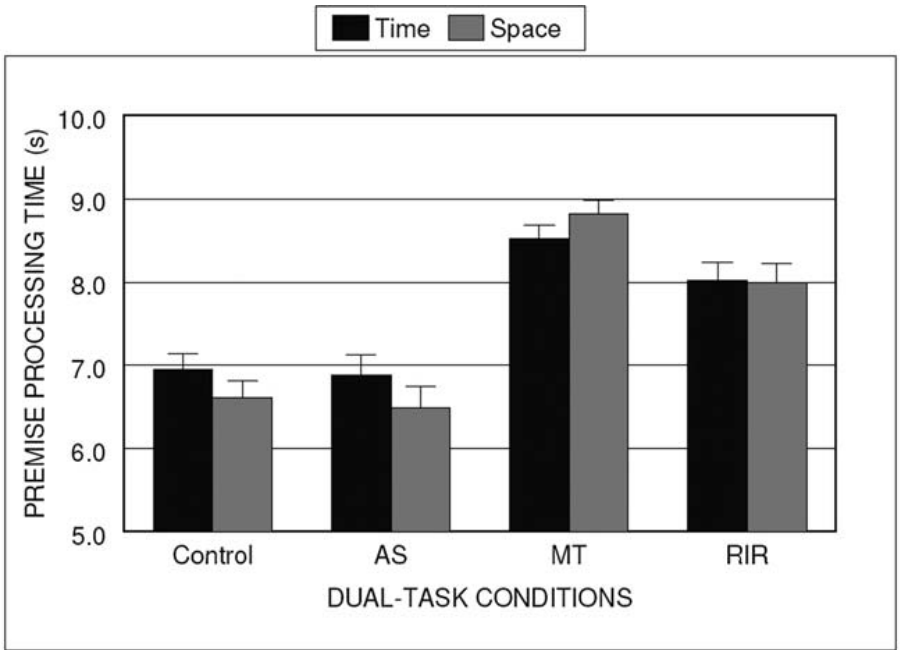


Fig. 1. Average premise processing times for problems with spatial and temporal content as a function dual-task conditions. The labels AS, MT and RIR refer to respectively articulatory suppression, matrix tapping and random interval repetition.

However, the most interesting aspect of these findings concerns the premise reading times in the self-paced condition. Figure 1 shows the average premise reading times as a function of dual task conditions and problem content (spatial or temporal relations). The figure clearly shows that compared to the control condition, only the matrix tapping and the random interval repetition conditions required more time to process the premises, while the articulatory suppression condition was not at all slowed down. So, it is clear that even though the premises are presented verbally, no verbal/phonological secondary task interference was observed. In other words, processing during premise intake mainly relied on visuo-spatial and executive processes.

Another interesting observation is that the delay due to matrix tapping is larger than the delay due to random interval repetition. It is clear from these data that this secondary task had a detrimental impact on the processing of the premise information, while it is known from previous research that it interferes with visuo-spatial rehearsal without requiring much attention (see e.g., Logie 1995, Smyth & Scholey 1992, 1994, 1996). Together

with the lack of articulatory interference this indicates that a visuo-spatial representation is built and maintained in visuo-spatial working memory.

2.2. The relational content

In the study just discussed, there are at least two reasons why the reasoners may have strongly relied on their visuo-spatial working memory resources. The first reason is that the problems were based on either spatial or temporal relations, which are easily mapped on a spatial representation, as already pointed out. The second reason is that problems could be one- or three-model problems. This variation in model complexity may have induced reasoners to rely more on a visuo-spatial representation.

It would be interesting to know, wether a visuo-spatial representation would be used when such a representation is not so strongly favoured by the conditions. To clarify this situation, other types of reasoning may be considered. Toms et al. (1993) worked with variations of conditional reasoning tasks. In such a problem, there are typically two premises. The first one expresses a conditional relation between two terms, such as `if it is a circle, then it is green`; the second premise expresses a simple statement, such as `it is green`. Over a series of studies in which Toms and colleagues varied the contents and the secondary tasks used, they found that conditional reasoning was impaired under a central executive load, but not with a concurrent visuo-spatial secondary task.

Klauer et al. (1997) used a more general variant of propositional reasoning and they found that when the problems expressed a spatial relationship, reasoning was impaired by a concurrent visuo-spatial task. Given that both in relational reasoning and in conditional reasoning based on spatial relations, visuo-spatial working memory seems to be implied, the hypothesis can be put forward that visuo-spatial representations are involved whenever the relations express a spatial order or position. In order to test this hypothesis Duyck et al. (2003) studied conditional reasoning about relations instead of entities. Two kinds of relations were studied in conditions with and without a concurrent spatial task (matrix tapping). One kind of relation was spatial, as in `if Pete lives to the right of Paul, then Stan does not live to the right of Kurt`. The other kind of relation was based on joint activities without explicit spatial order or location, as in `if Pete plays tennis with Paul, then Stan does not play tennis with Kurt`. Sixteen problems of each of these two kinds were constructed, and these were made as similar as possible except for the relation. These 16 problems were obtained as a factorial crossing of four problem types (*Modus ponens* or MP: $p \rightarrow q, p$; *Denial of the antecedent* or DA: $p \rightarrow q, \neg p$; *Affirmation of the consequent* or AC: $p \rightarrow q, q$; and *Modus tollens* or MT:

$p \rightarrow q, \neg q$) × four levels of negation in the first premise (both p and q positive, p negative and q positive, p positive and q negative, and both p and q negative). Half of the 42 participants were assigned to the condition with spatial problems and the other half solved the other (nonspatial) problems. The reading times needed for each of the two premises and for the solution were registered.

The most interesting part of the latency data concerns the reading times for the first premise. As can be seen in Figure 2, the concurrent spatial task did not slow down reading time for nonspatial problems (left panel), while it did for spatial ones (right panel). This effect was not moderated by problem types (MT, MP, ...), but it interacted with the presence of negations in the first premise. Actually the dual-task effect was much larger in

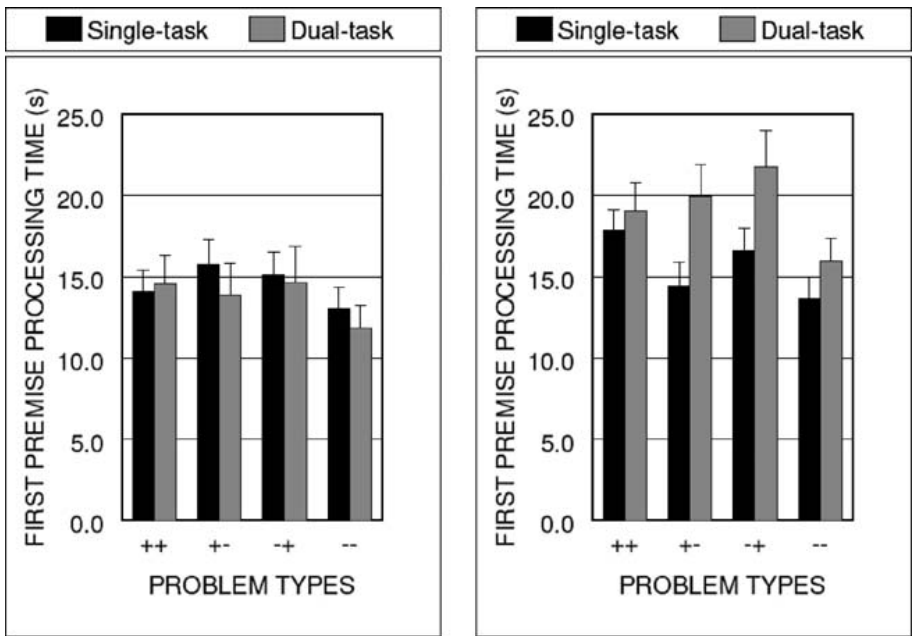


Fig. 2. Average premise processing times for the first premise on conditional reasoning problems as a function of the structure of negations in the first premise and task condition for nonspatial relations (left panel) and spatial relations (right panel).

problems with a negation than in those without negations.

In these findings, two aspects deserve to be discussed with respect to the role of the concurrent spatial task, namely its effect on spatial and nonspatial problems, on the one hand, and its effect on the presence of negations in the first premise. Concerning the first aspect, the present findings sug-

gest that the kind of relation, spatial rather than nonspatial, determines whether a problem will be represented with the help of visuo-spatial resources. Since the terms of the relations were the same in the two kinds of problems, only the kind of relation could affect the usage of visuo-spatial representations. There is no a priori reason to assume that the terms will be represented differently in the two kinds of problems. Interestingly, the findings are also consistent with those initially reported by Toms et al. (1993), because they used only relations of the nonspatial kind. Further research could show which other factors specifically determine the selection of the representational medium for the the models; it would, for example, be interesting to know whether the terms themselves play any role.

With respect to the presence of negations, it could be argued that a premise such as **if Pete lives to the right of Paul, then Stan does not live to the right of Kurt** will be translated into **if Pete lives to the right of Paul, then Stan lives to the left of Kurt**. It may be expected that such a transformation would cost time. However, the right panel of Figure 2 shows that, if anything, in the single-task conditions the premises with a negation are comprehended more quickly than those without a negation. What actually seems to happen is that in the presence of a concurrent spatial task, completing a representation of the premise information is slowed when it contains negations. If the premises would be converted as suggested above, there is no reason to expect that a spatial secondary task would affect this process. It seems more likely that first a spatial representation of the premise is built and that once this is completed a transformation is applied to this model when there is a negation. Because these operations are performed on the spatial representation, they are expected to be slowed down in the dual-task condition with a spatial load. For sure, this is not the end of the story. Future research could build on this finding to bring further clarification.

In summary, there seem to be two aspects that determine whether a relational problem will be represented by visuo-spatial models. The first is concerned with the structural complexity of the model(s). When the information cannot be completely represented by a simple string of tokens, the visuo-spatial modality may be used because of its power to represent structure. The second aspect relates to semantics. When the relation or the relational content has more affinity to a spatial representation, the visuo-spatial modality seems to be preferred, even though the reasoner remains in control of the choice actually made. This choice is not unlimited, but with verbally stated reasoning problems, the reasoner has the choice to build a verbal-phonological representation or a visuo-spatial representation. There are reasons to believe that the reasoner has access to information about the gains and the costs of each alternative (cf. models of strategical choice: Anderson 1990, Siegler & Shipley 1995).

3. Believability of relational models

In many studies of relational reasoning, the relations used are arbitrary ones, chosen in such a way that the reasoner *cannot rely on knowledge.* The reason for this is that knowledge may affect reasoning performance: a (tentative) conclusion which is at odds with our prior knowledge elicits more efforts to find counterexamples than a conclusion that is consistent with what we know. The result is that conclusions which are in agreement with our own beliefs are more often accepted than conclusions violating our beliefs. This is known as the *belief bias* effect. Most of the research on this effect has been performed in conditional and syllogistic reasoning (e.g., Evans et al. 2001, Klauer et al. 2000, Morley et al. 2004, Newstead et al. 1992, Oakhill & Johnson-Laird 1985, Quayle & Ball 2000, Revlin et al. 1980, Thompson et al. 2003). Although relational reasoning offers many interesting possibilities for such research because of the prominent reliance on visuo-spatial representations, research with knowledge-related materials has been rather scarce. The studies by Roberts & Sykes (2003) form a nice exception. They used premises with partly incorrect geographical or historical information, such as `The Pharaohs ruled after the Romans`; `The Pharaohs ruled before the Normans`; `The Pyramids were built at the time of the Romans`; `At the time of the Normans, William the Conqueror was king`; therefore `William the Conqueror was king after the Pyramids were built`. With this kind of problems, these investigators demonstrated a similar kind of belief bias effect as was observed in categorical reasoning (Evans et al. 1983, 1994, Newstead et al. 1992), namely that participants accepted believable conclusions more often than unbelievable ones and moreover within indeterminate problems, the difference was larger for invalid than for valid conclusions.

Considered within the perspective of the mental models theory, not only the conclusions can match or contradict general knowledge. It is also possible that the premises themselves describe situations (models) that do or do not correspond with general knowledge. A context in which such effects may typically occur is related to scripted activities. Within the context of schema theory, scripts are considered as a particular kind of cognitive schema that can be described as a network consisting of nodes and (mostly temporal) relations between the nodes. Some nodes are fixed, others are variable or optional and can be filled in when the script is activated and for some of these variable and optional nodes the script may provide defaults that can be overridden (see e.g., Abelson 1981, Bower et al. 1979, Brewer & Treyens 1981, Graesser & Nakamura 1982, Nakamura et al. 1985). Typical scripts are a visit to a restaurant, to the library, to a doctor, to a dentist, etc. Besides these temporally ordered scripts, there exist also unordered

scripts, such as the circus script which specifies a number of acts typically performed in a circus, but the order of the acts is not determinate.

How scripts can be used to manipulate the believability of the entire model given by the premises is illustrated by a study of Dierckx et al. (2004). They used ordered scripts to create believable and unbelievable problems, which were then compared with neutral problems based on unordered scripts. Neutral problems were included in order to allow inferences on whether facilitation (better than neutral) or suppression (worse than neutral) was at the basis of the observed differences between believable and unbelievable problems. For the construction of believable and unbelievable problems an ordered script such as a visit to the dentist can be used. A typical sequence of events for this script is: **open mouth**, **localise toothache**, **anaesthetize**, **swallow saliva**, and **plug tooth**. Table 1 shows how these activities can be combined in a series of all believable premises that taken together either add up to a believable (left panel) or an unbelievable model (central panel). The table also shows how an unordered script (circus) is used to create neutral problems.

The most difficult part concerns the ordered scripts. In the believable problems, for example, four of the five events mentioned form a strict ordering; one event (**swallow saliva**) is a typical script event that can occur anywhere in the sequence. The usage of such an event in each script, makes it possible to create premises that are believable while the entire sequence may be unbelievable because it violates the normal script order. The central panel of the table shows how this is realized. Interestingly, the validity of the proposed conclusion can be varied independently from the variation in believability, as is shown in the bottom panels of Table 1.

Dierckx et al. (2004) used problems like these to study model believability effects. For half of the participants a header referring to the script was added. The authors found effects of model believability for invalid but not for valid conclusions, as is shown in Figure 3. The solution latencies (left panel) of invalid conclusions were faster in believable than in unbelievable problems, faster in believable than in neutral problems, but not faster in neutral than in unbelievable problems. In comparison to the neutral problems, accuracy (right panel) was highest for the believable problems and lowest in the unbelievable problems and both were different from the neutral. This shows that evocation of the script supports reasoning when the model is believable and interferes when reasoning with the model is not believable. It should be noted that this comparison is only based on conclusions that were not contaminated by possible other effects: The same relations between the second and the fourth term was tested in all conditions. Interestingly and as expected, the facilitating effect of believable problems was not only present in solution time and accuracy; it was also observed already during premise reading. Moreover, this study reveals an

Table 1
Examples of the construction of a believable, an unbelievable and a neutral model from sets of believable premises based on script events in the study of Dierckx et al. (2004). The neutral problem is based on an "unordered" script and the actions A-E are presented in any order.

Script actions		
A: open mouth		A: tame lions
B: localise toothache		B: ride small bicycle
C: anaesthetize		C: throw cones
X: swallow saliva		D: perform clown act
D: plug tooth		E: train horses
Example 1	Example 2	Example 3
Believable problem	Unbelievable problem	Neutral problem
A before B	A before B	A before B
B before C	B before D	B before C
C before X	D before X	C before D
X before D	X before C	D before E
Model representation		
A-B-C-X-D	A-B-D-X-C	A-B-C-D-E
Valid conclusions		
B-X	B-X	B-D
Invalid conclusions		
X-B	X-B	D-B

effect of believability (of the set of premises) while the believability of the individual premises and the conclusion was held constant. This effect is distinct from the believability as studied thus far in the literature, namely, the effect of believability of the premises themselves and of the believability of the conclusion.

By and large, these findings are consistent with the view that during premise presentation a script is triggered and helps to maintain a consistent (believable) model, but seems to interfere for the maintenance of (unbelievable) or inconsistent models. The main question is, whether these effects are mediated by the interaction of long-term and working memory. As a further test of this hypothesis, in an unpublished study, half of the participants were required to do some calculations after premise presenta-

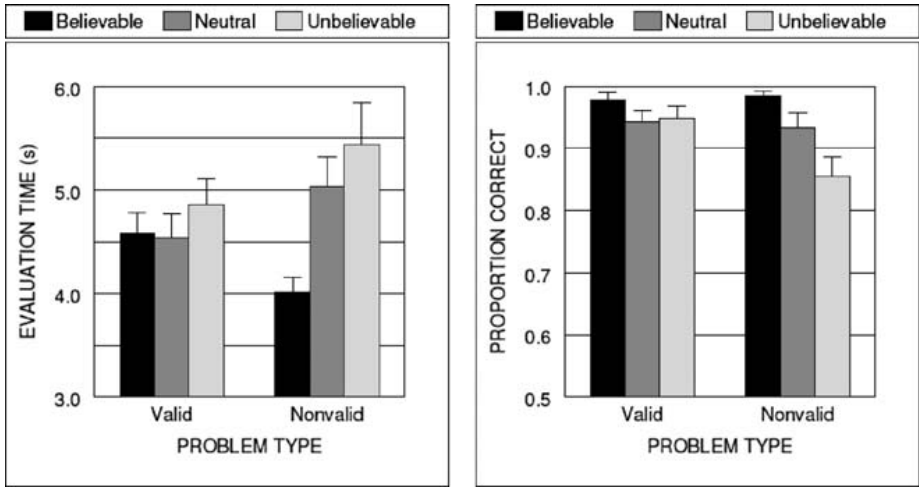


Fig. 3. Average solution latency (left) and proportion of correct solutions (right) as a function of model believability and conclusion validity in the one-model script-based problems of Dierckx et al. (2004).

tion and before conclusion verification. If working memory is used to build and maintain a temporary model of the premise information, it may be expected that the extra task which requires both storage and processing of information will compete with the working memory resources needed for the maintenance of the model. Hence, the differences between believable and unbelievable models should be enhanced.

As expected, the additional calculation task had no effect on the verification accuracy of valid conclusions; it also had no effect on the accuracy of invalid conclusions in believable and neutral problems, but it dramatically lowered the accuracy of invalid unbelievable conclusions. This effect is shown in Figure 4. Solution latencies were not affected much by the additional arithmetic task, because in all cases, the model has been constructed by the end of the premise presentation phase. It is important to point out that in this study, a working memory load was used which interfered with the executive control processes, without itself requiring any storage. Since the neutral problems were basically not affected by this load, it is clear that the effect is not due to competition for temporary storage. To the contrary, the dramatic performance decrease in the unbelievable problems must be accounted for by an increased competition between the model maintained in working memory and the activated script in long-term memory. Because of the load on executive control created by the simple arithmetic task (see e.g., Deschuyteneer & Vandierendonck 2005), insufficient executive resources were left for an appropriate control of the interference between the conflicting working memory and long-term memory representations.

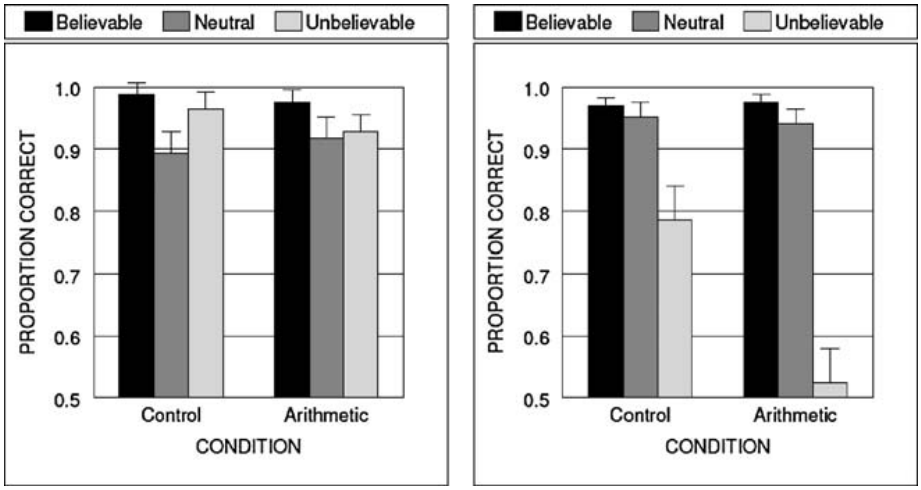


Fig. 4. Average accuracy of correct solutions in valid (left) and invalid (right) conclusion verifications as a function of experimental condition and model believability in one-model script-based problems.

This set of findings with respect to model believability shows that prior knowledge may facilitate or hamper reasoning performance. In particular, the activation of a model imbedded in knowledge (a script) which is part of semantic memory, may help reasoning if this model matches the information given in the premises. This is the case for premises resulting in a believable model. In contrast, when the model imbedded in the script activated in semantic memory does not correspond exactly to the information in the premises, then the model specified by the premises is unbelievable and this results in both slower and less accurate reasoning. It must be stressed, though, that both the positive effect of believable models and the negative effect of unbelievable models is not a general characteristic. In fact, these effects do only seem to occur essentially when the conclusion to be drawn is invalid. For valid conclusions, believability does not seem to matter so much. This could be an artefact, however, because valid conclusions in unbelievable problems do not contradict the script (prior knowledge) and so the effect of believability may not be playing any role in such conclusions.

Another aspect that deserves attention is that in all experiments discussed, attempts were made to include a neutral condition. Although this is not easy to realize, the usage of unordered scripts seems to yield an appropriate procedure for the construction of such neutral problems.

There is much room for strengthening and elaborating these findings. Nevertheless, with the data available it seems that prior knowledge stored as scripts may be activated by the problem context and is then present while

the model derived from the premises is constructed. When the constructed model and the activated model match, the task for the reasoner becomes easier because all what is needed is that the script is kept active. On the contrary, when the constructed and the activated model differ, additional (executive) control processes, such as conflict monitoring, will be needed to suppress the activated model and to maintain the newly constructed model. Because this operation consumes more resources, reasoning should become slower and more error-prone and this effect should be enhanced when the resources are depleted by additional processes. This was confirmed in the study with an intervening arithmetic task and it resulted in larger effects of model believability.

4. Specificity of relational information

Thus far, the present chapter focused on one particular kind of long-term memory effects accounted for by contextual information. As shown, relevant background knowledge may help or impair the reasoning performance. There are, however, other possibilities to study the interaction of long-term and working memory. In genealogical reasoning, for example, the relation itself allows for a tight coupling between working memory and long-term memory, because kinship information is given directly via memory representations (Miller & Johnson-Laird 1976, sec. 5.2). Kinship relations are abstract concepts; they cannot be characterized in terms of obvious external physical characteristics. There is no way one can tell from perceptual evidence alone whether Paul is Billy's cousin. Kinship terms have no "concrete" referents. In fact, they are purely cultural in content. Even a concept like "father" may have different meanings in different cultures; in Western Europe, it may refer to the biological relation (`Albert is the biological father of Delphine`) or to an adoptive parenthood (`Mike is the adoptive father of Lucia`). Moreover, different societies have very different ways of categorizing relatives; the Western European system is only one of many. For example, in Chinese (Mandarin) older brother and younger brother are separate terms (i.e., "gege" and "didi") and in the Samoan kinship terminology, siblings and cousins may be referred to by the same kin terms, e.g. the term "tuafafine" may refer to the sister of a male anchor, as well as the female cousin of a male anchor (Jonsson 1999). Kinship relations seem ideally suited to study the interaction of knowledge and working memory in reasoning about relations. An additional advantage is that kinship terminology represents a fairly compact, well-defined set of relations which, while being small enough to handle easily, also varies considerably in complexity compared to the rather sim-

ple relations traditionally used in relational reasoning tasks (e.g., "taller than"), spatial relations (e.g., "on the left of") or temporal relations (e.g., "before"). Relations such as "father of," "uncle of," and "grandchild of" are intransitive, while the "ancestor of" and "descendant of"-relations are transitive.

Family relations are often displayed in a spatial format, with vertically the descendency relations (grandfather, father, grandmother, mother, grandparent, parent, son, grandson, daughter, granddaughter, child, grandchild, ...) and with horizontally the sibling relations (brother, sister) and the marriage relations. People have learned to use the very specific labels representing different types of relationships as well as the spatial representation of the family tree (see e.g., Wood & Shotter 1973). On the basis of this, one may expect that people have acquired a number of logical rules to make inferences from given family relations, but they will also have learned to build (spatial) models to represent specific situations.

A first question, then, concerns the issue whether people indeed have a preference for the usage of mental models to represent given information as is apparent for other types of relational reasoning. There are reasons to believe that this is indeed the case, but the present scope prohibits development of these arguments (but see Dekeyser 1997, for more information on this issue). Therefore, we shall assume that people indeed solve genealogical reasoning questions by means of the representation of mental models.

The second question, concerns the issue whether the mental models constructed on the basis of genealogical relations are in any important way different from the mental models constructed from other order relations. In a typical relational reasoning problem, reasoners are given order information specifying for example that someone is older than someone else. The spatial representation is sufficiently detailed when the terms (tokens) are given a spatial position that corresponds to the order as expressed in the premises. If `Ned is older than Olivia` and `Olivia is older than Peter`, then it does not matter whether `Olivia` is only a little bit or much older than `Peter`. All what is needed to infer the correct conclusion that `Ned is older than Peter` is the order of the terms. In genealogical reasoning, care must be taken that also the genealogical distance is represented when it is needed. If `Roger is the grandfather of Stephen` and `Stephen is the father of Tom`, the inferred relationship `Roger is the great-grandfather of Tom`, must include the distance between the terms. One possible solution would be to make explicit that `Roger is the father of some x` and that this `x is the father of Stephen` and to build a representation that explicitly states this. In other words, the two premises would result in the model "`Roger - x - Stephen - Tom`" from which it can be inferred that `Roger` comes three steps before `Tom`, so that `Roger is the great-grandfather of Tom`. Making such explicit repre-

sentations of the relations costs quite some effort in addition to the need for representational resources. Hence, it may be expected that people will be reluctant to construct such explicit models. Nevertheless, the reasoner must take care to develop a correct representation. This could be achieved in several ways. Firstly, it is possible to place markers that indicate that the relationship can be fleshed out. Where the x-elaboration discussed above gives a precise specification of the distance, it may also be possible to insert a marker that there is a non-standard distance (e.g., `Roger -()- Stephen`), so that later on, if needed an elaboration is possible by specifying `Roger - x - Stephen` for the grandfather relation or `Roger - x - y - Stephen` for expressing a great-grandfather relation. Given some ideas present in the mental models theory and its developments, one could expect that people would follow this strategy (e.g., annotations: Vandierendonck et al. 2004). Another possibility is that reasoners represent the distance without adding any tokens. For the two premises in the last example, this would result in a representation such as "`Roger --- Stephen - Tom`". A further possibility is that reasoners will only represent the additional distance information if they expect to need it. In a context of problems about who is who's ancestor, they will probably suffice with the typical relational model, as the inference `Roger is an ancestor of Tom` does not call on the distance information at all.

4.1. Study 1: Transitive versus intransitive inference

In order to clarify how the relations are represented in different kinds of reasoning contexts, we compared reasoning with order and genealogical relations (Van der Beken & Vandierendonck 2005). This was realized in three different conditions. In a first condition, reasoners were given four premises based on the transitive relation "older than" (for an example, see left panel of Table 2). After reading the premises, the reasoners were asked to verify which one of two statements expressing the relation between the second and the fourth term was correct. In the other conditions, similar premises were presented with the relation "father of" (see middle and right panel of Table 2). In the second condition, the reasoners were asked to verify the "ancestor" relationship between the second and the fourth terms. In the third condition, a genealogical relation had to be verified between the two terms.

Figure 5 displays the main findings with respect to the solution latencies. A first observation is that the solution times for the intransitive problems were longer than for the transitive problems, irrespective of whether the latter were based on temporal or on genealogical premises. Secondly, in both conditions where transitive conclusions were verified, the typical *dis-*

Table 2
Examples of the three types of problems used in the comparison of reasoning between transitive and genealogical relations.

Condition 1	Condition 2	Condition 3
Relational	Temporal	Genealogical
Transitive	Genealogical	Intransitive
Example 1	Example 2	Example 3
Believable problem	Unbelievable problem	Neutral problem
A before B	A before B	A before B
B before C	B before D	B before C
C before X	D before X	C before D
X before D	X before C	D before E
Linear	Genealogical	
Peter older than Roger	Peter father of Roger	
Roger older than Klaus	Roger father of Klaus	
Klaus older than Steve	Klaus father of Steve	
Steve older than Willy	Steve father of Willy	
Transitive	Transitive	Intransitive
Roger older than Steve	Roger ancestor of Steve	Roger grandfather of Steve
Steve older than Roger	Steve ancestor of Roger	Steve grandfather of Roger
		Neither of both[a]

[a] In the conditions with transitive relations, the alternative "neither of both" was not used because this is not a possible alternative. In a control experiment, the number of alternatives was equal across conditions, but that did not change the findings.

tance effect was observed: the further apart the elements in the unified representation the faster the inference was made.[4] In the condition with verification of intransitive inferences, however, a *reversed distance effect*

[4] Figure 5 does not show this very clearly; the figure suggests no difference; nevertheless, the effect was statistically reliable.

was observed, in such a way that the longer the distance between the elements in the verification, the longer it took to decide on the correct answer. While the distance effect is generally taken as support for the hypothesis that an integrated spatial-array based representation of the premise information has been constructed (e.g., Maybery et al. 1986, Potts 1974, 1976), the reversed distance effect has been found in situations represented also spatially but requiring a more precise inference (e.g., Mayer 1978, 1979).

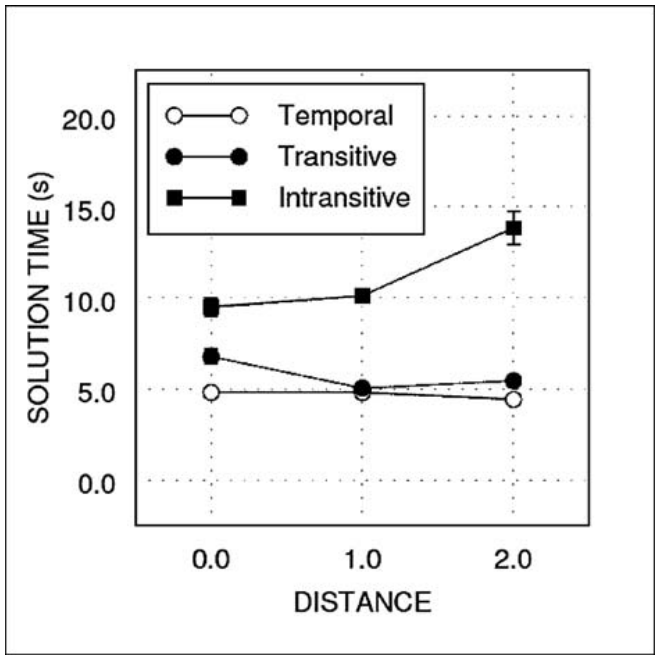


Fig. 5. Average solution latency of correct solutions (standard errors in the bars) as a function of problem type and distance between the queried terms in linear and genealogical reasoning problems.

In the accuracy data, the differences were rather small and not significant. The pattern of findings is completely consistent with the idea that for all types of problems a spatial model is constructed in which only the order of the tokens is taken into account. For the transitive problems, this information suffices to easily infer whether a particular term precedes another one in the array. However, for the intransitive problems, extra processing is required to convert the distance into an exactly labelled relationship. For a distance-0 relation, the relation must not be renamed, but for distance-1 (e.g., grandfather) and distance-2 (e.g., great-grandfather) the number of intervals must be counted to make sure that the relation between the

Table 3
Examples of the three types of problems used in the comparison of reasoning between genealogical problems with and without second-order relations.

Condition 1	Condition 2	Condition 3
Only first-order	Early second-order	Late second-order
Peter father of Roger	Peter father of Roger	Peter father of Roger
Roger father of Klaus	Roger grandfather of Klaus	Roger father of Klaus
Klaus father of Steve	Klaus father of Steve	Klaus father of Steve
Steve father of Willy	Steve father of Willy	Steve grandfather of Willy
Transitive inference		
Roger is ancestor of Steve		
Steve is ancestor of Roger		
Intransitive inference		
Roger is great-grandfather of Steve		
Steve is great-grandfather of Roger		

two terms exactly matches the relational term specified in the proposed conclusion.

4.2. Study 2: Representation of second-order relations

In the study just discussed, all the premises contained only direct relations so that for the genealogical problems there was no need to represent differences in distance between particular terms. In a second study, second-order relations (such as grandfather) were introduced. There were three conditions: only first-order relations, one second-order relation between the second and the third term (second premise) in the linear array and one second-order relation between the fourth and the fifth term (fourth premise). The design is shown by means of examples in Table 3.

The results of this 3 (problem type) × 2 (inference type) design are displayed in Figure 6 for the solution latencies with in the left panel the results for the transitive problems and in the right panel the results for the intransitive problems. Whether the relations are first-order or second-order should not matter much for transitive problems as only the order of two terms in the spatial array must be verified. Hence, a standard dis-

tance effect was predicted for these problems irrespective of the presence or the location of the second-order relations. This expectation was confirmed: There was no performance difference between the three problem types and, in all three cases, the (normal) distance effect showed up. For the intransitive problems, on the contrary, it was expected that the introduction of second-order relations would reverse the distance effect because most of the problems now contain relations which have to be represented differently in the model. More specifically, we expected a strong reversed distance effect for the problems with a second-order relation early in the problem, because the additional information was necessary and had to be processed to obtain the correct relational distance between the terms. For the problems containing only first-order relations and for the problems with a late second-order relation, on the contrary, a rather small reversed distance effect was expected, because the second-order relation was never involved in the target problems. In this case, the effect was expected to be small because the constructed model may contain additional relational information, but since it is of no use for the solution of the problem, the reversed distance effect was expected to be rather weak. Figure 6 shows that these expectations were completely confirmed.

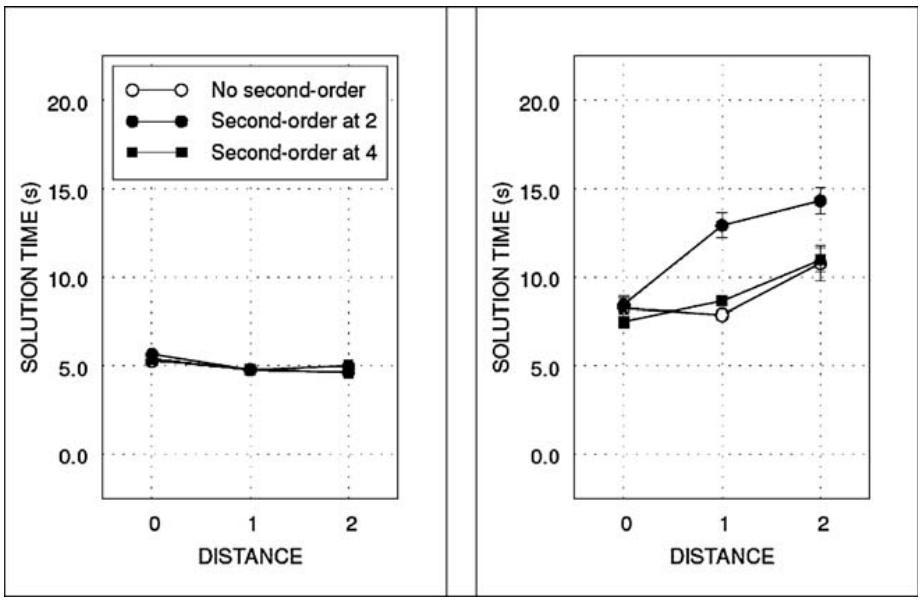


Fig. 6. Average solution latencies (standard errors in the bars) for transitive (left) and intransitive genealogical problems (right panel) as a function of problem type and distance between the queried terms.

The study of reasoning with kinship relations is rather new. Much remains to be discovered. However, the data reviewed here, show already that (a) linear reasoning with kinship is different from relational reasoning as it is usually studied because kinship relations are not transitive. Interestingly, instead of a distance effect a reversed distance effect is observed. A second element is that the semantic basis and elaboration of the kinship relations also plays a role. While this aspect was not strongly evident in the research results presented here, it is possible that people use their expertise regarding their own families in solving genealogical problems, especially when representing kinship relations with more than one basic definition. As pointed out by Oden & Lopes (1980, 489), a person who was raised in close contact with a large extended family, may see "cousins" as "the other grandchildren of their grandparents," while another person with less intense family contacts may see "cousins" as "children of aunts and uncles." Exactly because the domain of kinship terms is so semantically rich and variable, study of reasoning about genealogical relations provides a very interesting domain for testing processing theories about deductive reasoning.

5. Summary and conclusions

The findings reported in the present chapter further add to the accumulating body of evidence supporting the mental models theory as an adequate account of reasoning with relational information. By and large, the findings presented corroborate the hypothesis that premise information is used to construct a representation in visuo-spatial working memory. The main characteristics of such a representation or model are (a) that it is transient and (b) that the relations are represented spatially. The first characteristic indicates that the model is maintained only long enough to infer a conclusion from it and that it is discarded very soon after reaching the conclusion. The second characteristic refers to the fact that any relation can be represented spatially, even though this is not equally easy for all types of relations. It is also clear from the literature on relational reasoning that the representation constructed is minimal. If there is no need, no detailed information is maintained and most often only the order of the terms in the representation is modelled.

Whereas relational reasoning seems to be based on spatial models and such modeling is available for all kinds of relations, it seems that reasoning about relations does not always depend on the visuo-spatial modality. The usage of visuo-spatial working memory seems to be preferred in situations where the representation of the problem information is rather complex (not just linear series) and in situations where the relations can be easily

mapped on a spatial representation. This is clearly observed in problems where the number of models required to attain the solution varies so that at least for an important subset of the problems a more complex representation is required (e.g., Vandierendonck et al. 2004). When the relation is spatial or easily mapped on a spatial relation, then reasoners will also preferentially call on the visuo-spatial working memory component. This was, for example, the case in conditional reasoning about spatial relations, but not in conditional reasoning about nonspatial relations (Duyck et al. 2003). This does not mean that a spatial representation is obligatory in such circumstances, but rather that it is the easiest route to a solution. We suggest, however, that reasoners have control over which representational medium they use for the representation of the models and over what kind of information is maintained in working memory in the service of finding a solution. There is no direct evidence available on these issues, but the findings of Dierckx et al. (2003) suggest that strategic control over the kind of information that is entered into the models is a genuine possibility.

The present chapter has shown that the model constructed in working memory is based on all the information available. In particular, if the premises are reminiscent of a particular known situation (such as a script), then this script seems to be activated. If the script and the premises match, reasoning performance is good, but if the script and the premises contradict each other on certain points, reasoning is impaired. Interestingly, this believability effect is entirely due to the model (or the script) and not to the premises as they all were individually believable, nor to the believability of the conclusions because these were all equally believable. It was also shown that the maintenance of an unbelievable model suffered from an extra load on working memory. This shows that reasoners must do more effort to maintain a model that is inconsistent with their knowledge than to maintain a model that is completely consistent or that is simply possible. These effects essentially show up in the verification of invalid conclusions. The latter kind of conclusions are inconsistent with the model, but they may be consistent with the script. Hence, knowledge about the world seems to be activated by reading the premises and reasoners cannot escape using this knowledge even if it does not match the information specified in the premises. Together with recent findings about knowledge effects in conditional reasoning (e.g., De Neys et al. 2002, 2003*a*,*b*, Markovits & Potvin 2001, Quinn & Markovits 1998), this supports the main thesis of the mental models theory, namely that reasoning is a semantically based process.

The usage of genealogical relations seems to open new avenues for the study of relational reasoning. In this chapter a few studies were briey discussed. These studies show that basic genealogical relations (such as "father of," "son of") are represented in the same way as the standard relations used in many studies. This representation seems to be restricted to an or-

dering of the tokens or terms involved in the premises. When the solution can be expressed in the same type of relation as in transitive problems, the traditional distance effect is obtained, but when an intransitive relation must be verified, extra processing seems to be required, resulting in slower solution, more errors and a reversed distance effect in latencies. Furthermore, when the premises also contain second-order relations, then the model constructed seems to change. As long as the reasoners expect transitive problems, the modeling seems to be minimal and a standard distance effect is observed. However, if the situation is such that an important subset of the problems is intransitive, then the distance effect is reversed and this reversal is extremely strong when the second-order relation is relevant for the verification of the problem.

Taken altogether, the series of studies reviewed in this chapter, support the view that there is *a tight coupling between working memory and long-term memory.* Believable models, or, in other words, models that have a pre-existing representation in long-term memory seem to be more easy to construct and these representations seem to take some precedence over the models constructed in working memory.[5] Likewise, when particular representation formats in long-term memory are activated by the premises as with spatial relations, then visuo-spatial working seems to be used as the platform for constructing the model, even when this is normally not the case (as in conditional reasoning). Finally, the work on genealogical reasoning suggests that the long-term labels for relations are used flexibly. Only as much information seems to be used as is necessary for handling the problems. In other words, working memory seems to be at the service of reasoning and is used in a flexible way depending on the goals of the reasoner and the constraints imposed by the task. This kind of interaction between long-term and working memory is probably the rule rather than the exception. Studies such as the ones discussed here play a crucial role in this approach: At the same time they allow to augment our knowledge about reasoning and about working memory while elucidating how these two memory components collaborate in the service of cognitive tasks. If Baddeley's hypothesis of an episodic buffer is correct (Baddeley 2000), it is quite likely to appear in tasks and task situations as the ones discussed here. Thus far, not much direct evidence has been collected to support the concept of episodic buffer. However, there are many reasons to assume that somehow information in long-term memory and in working memory

[5] An alternative interpretation could be that knowledge is used as a back-up when maintenance of the model in memory fails. However, such a view has difficulty in explaining the observation that the interference already occurs during premise reading. It would be interesting, therefore, to see more research that explores how knowledge affects the construction of models.

must interact in different tasks. The present chapter has shown a number of ways in which such an interaction may occur without specifying exactly the mechanism which is at the basis of this interaction. Given that the reasoning problems considered here are always given in a verbal format and that to some extent visuo-spatial working memory seems to be involved in solving these problems, it is clear that when considering only the representations in working memory there must exist some way to relate the verbal (phonological?) and the visuo-spatial representations. If there would be clear evidence for the existence of the episodic buffer as a mechanism to support these interactions within working memory, the same device would, no doubt, also be useful as a vehicle for the interactions between working memory and long-term memory. For that reason, it might be worthwhile to start to develop methods to specify the mechanism underlying the interactions between long-term and working memory.

References

Abelson, R. P. (1981), 'Psychological status of the script concept', *American Psychologist* **36**, 715–729.

Anderson, J. R. (1990), *The Adaptive Character of Thought*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Baddeley, A. (1986), *Working Memory*, Oxford University Press, Oxford.

Baddeley, A. (2000), 'The episodic buffer: A new component of working memory?', *Trends in Cognitive Sciences* **4**, 417–423.

Baddeley, A. D. & Hitch, G. (1974), Working memory, *in* G. H. Bower, ed., 'The Psychology of Learning and Motivation', Vol. 8, Academic Press, New York, pp. 47–89.

Baddeley, A. D., Lewis, V. J. & Vallar, G. (1984), 'Exploring the articulatory loop', *Quarterly Journal of Experimental Psychology* **36A**, 233–252.

Baddeley, A. D. & Logie, R. H. (1999), Working memory: The multiple-component model, *in* A. Miyake & P. Shah, eds, 'Models of Working Memory: Mechanisms of Active Maintenance and Executive Control', Cambridge University Press, Cambridge, pp. 28–61.

Barclay, J. R. (1973), 'The role of comprehension in remembering sentences', *Cognitive Psychology* **4**, 229–254.

Bower, G. H., Black, J. B. & Turner, T. J. (1979), 'Scripts in memory for text', *Cognitive Psychology* **11**, 177–220.

Brewer, W. F. & Treyens, J. C. (1981), 'Role of schemata in memory for places', *Cognitive Psychology* **13**, 207–230.

Byrne, R. M. J. (1989), 'Suppressing valid inferences with conditionals', *Cognition* **31**, 61–83.

Byrne, R. M. J., Espino, O. & Santamaria, C. (2000), Counterexample availability, *in* W. Schaeken, G. D. Vooght, A. Vandierendonck & G. d'Ydewalle, eds,

'Deductive Reasoning and Strategies', Lawrence Erlbaum Associates, Mahwah, NJ, pp. 97–101.

Clark, H. H. (1969), 'Linguistic processes in deductive reasoning', *Journal of Experimental Psychology* **76**, 387–404.

Clark, H. H. (1971), 'More about "adjectives, comparatives, and syllogisms"', *Psychological Review* **78**, 505–514.

De Neys, W., Schaeken, W. & d'Ydewalle, G. (2002), 'Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework', *Memory & Cognition* **30**, 908–920.

De Neys, W., Schaeken, W. & d'Ydewalle, G. (2003*a*), 'Causal conditional reasoning and strength of association: The disabling condition case', *European Journal of Cognitive Psychology* **15**, 161–176.

De Neys, W., Schaeken, W. & d'Ydewalle, G. (2003*b*), 'Inference suppression and semantic memory retrieval: Every counterexample counts', *Memory & Cognition* **31**, 581–595.

De Soto, C. B., London, M. & Handel, S. (1965), 'Social reasoning and spatial paralogic', *Journal of Personality and Social Psychology* **4**, 513–521.

Dekeyser, M. (1997), Genealogisch redeneren: explicitatie en toetsing van de mentale model opvatting [genealogical reasoning: Explicitation and examination of the mental model account], Master's thesis, University of Leuven.

Deschuyteneer, M. & Vandierendonck, A. (2005), 'Are 'input monitoring' and 'response selection' involved in solving simple mental additions', *European Journal of Cognitive Psychology* **17**, 343–370.

Dierckx, V., Vandierendonck, A., Liefooghe, B. & Christiaens, E. (2004), 'Plugging a tooth before anaesthetising the patient? The influence of people's beliefs on reasoning about the temporal order of actions', *Thinking and Reasoning* **10**, 371–404.

Dierckx, V., Vandierendonck, A. & Pandelaere, M. (2003), 'Is model construction open to strategic decisions? An exploration in the field of linear reasoning', *Thinking and Reasoning* **9**, 97–131.

Duyck, W., Vandierendonck, A. & De Vooght, G. (2003), 'Conditional reasoning with a spatial content requires visuo-spatial working memory', *Thinking and Reasoning* **9**, 267–287.

Evans, J. S. B. T. (2000), What could and could not be a strategy in reasoning, *in* W. Schaeken, G. D. Vooght, A. Vandierendonck & G. d'Ydewalle, eds, 'Deductive Reasoning and Strategies', Lawrence Erlbaum Associates, New York, pp. 1–22.

Evans, J. S. B. T., Barston, J. L. & Pollard, P. (1983), 'On the conflict between logic and belief in syllogistic reasoning', *Memory & Cognition* **11**, 295–306.

Evans, J. S. B. T., Handley, S. J. & Harper, C. N. J. (2001), 'Necessity, possibility and belief: A study of syllogistic reasoning', *Quarterly Journal of Experimental Psychology* **54A**, 935–958.

Evans, J. S. B. T., Newstead, S. E., Allen, J. L. & Pollard, P. (1994), 'Debiasing by instruction: The case of belief-bias', *European Journal of Cognitive Psychology* **6**, 263–285.

Evans, J. S. B. T., Newstead, S. E. & Byrne, R. M. J. (1993), *Human Reasoning:*

The Psychology of Deduction, Lawrence Erlbaum Associates, Hillsdale, NJ.

Foos, P. W., Smith, K. H., Sabol, M. A. & Mynatt, B. T. (1976), 'Constructive processes in simple linear-order problems', *Journal of Experimental Psychology: Human Learning and Memory* **2**, 759–766.

Gilhooly, K. J., Logie, R. H., Wetherick, N. E. & Wynn, V. (1993), 'Working memory and strategies in syllogistic reasoning tasks', *Memory & Cognition* **21**, 115–124.

Gilhooly, K. J., Logie, R. H. & Wynn, V. (1999), 'Syllogistic reasoning tasks, working memory, and skill', *European Journal of Cognitive Psychology* **11**, 473–498.

Goodwin, G. P. & Johnson-Laird, P. N. (2005), 'Reasoning about relations', *Psychological Review* **112**, 468–493.

Graesser, A. C. & Nakamura, G. V. (1982), The impact of schema on comprehension and memory, *in* G. H. Bower, ed., 'The Psychology of Learning and Motivation', Vol. 16, Academic Press, New York, pp. 59–109.

Henle, M. (1962), 'The relation between logic and thinking', *Psychological Review* **69**, 366–378.

Huttenlocher, J. (1968), 'Constructing spatial images: A strategy in reasoning', *Psychological Review* **75**, 550–560.

Johnson-Laird, P. N. (1983), *Mental Models*, Cambridge University Press, Cambridge.

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

Johnson-Laird, P. N. & Byrne, R. M. J. (1989), 'Only reasoning', *Journal of Memory and Language* **28**, 313–330.

Jonsson, N. (1999), Some Grammatical Properties of Samoan Kin Terms, Master's thesis, Stockholm University, Department of Linguistics.
URL: *http://www.ling.su.se/staff/niki/writings/501ver-d.pdf*

Klauer, K. C., Musch, J. & Naumer, B. (2000), 'On belief bias in syllogistic reasoning', *Psychological Review* **107**, 852–884.

Klauer, K. C., Stegmaier, R. & Meiser, T. (1997), 'Working memory involvement in propositional and spatial reasoning', *Thinking and Reasoning* **3**, 9–47.

Logie, R. H. (1995), *Visuo-Spatial Working Memory*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Markovits, H. & Potvin, F. (2001), 'Suppression of valid inferences and knowledge structures: The curious effect of producing alternative antecedents on reasoning with causal conditionals', *Memory & Cognition* **29**, 736–744.

Maybery, M. T., Bain, J. D. & Halford, G. S. (1986), 'Information-processing demands of transitive inference', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **12**, 600–613.

Mayer, R. E. (1978), 'Qualitatively different storage and processing strategies used for linear reasoning tasks due to meaningfulness of premises', *Journal of Experimental Psychology: Human Learning and Memory* **4**, 5–18.

Mayer, R. E. (1979), 'Qualitatively different encoding strategies for linear reasoning premises: Evidence for single association and distance theories', *Journal of Experimental Psychology: Human Learning and Memory* **5**, 1–10.

Meiser, T., Klauer, K. C. & Naumer, B. (2001), 'Propositional reasoning and working memory: The role of prior training and pragmatic content', *Acta Psychologica* **106**, 303–327.

Miller, G. A. & Johnson-Laird, P. N. (1976), *Language and Perception*, Cambridge University Press, Cambridge, MA.

Miyake, A. & Shah, P., eds (1999), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, Cambridge University Press, Cambridge.

Morley, N. J., Evans, J. S. B. T. & Handley, S. J. (2004), 'Belief bias and figural bias in syllogistic reasoning', *Quarterly Journal of Experimental Psychology* **57A**, 666–692.

Mynatt, B. T. & Smith, K. H. (1977), 'Constructive processes in linear order problems revealed by sentence study times', *Journal of Experimental Psychology: Human Learning and Memory* **3**, 357–374.

Nakamura, G. V., Graesser, A. C., Zimmerman, J. A. & Riha, J. (1985), 'Script processing in a natural situation', *Memory & Cognition* **13**, 140–144.

Newstead, S. E., Pollard, P., Evans, J. S. B. T. & Allen, J. L. (1992), 'The source of belief bias in syllogistic reasoning', *Cognition* **45**, 257–284.

Norman, D. A. & Shallice, T. (1986), Attention to action: Willed and automatic control of behavior, *in* R. J. Davidson, G. E. Schwarts & D. Shapiro, eds, 'Consciousness and Self-Regulation', Vol. 4, Plenum Press, New York, pp. 1–18.

Oakhill, J. V. & Johnson-Laird, P. N. (1985), 'The effects of belief on the spontaneous production of syllogistic conclusions', *Quarterly Journal of Experimental Psychology* **37A**, 553–569.

Oden, G. C. & Lopes, L. L. (1980), 'Kin search: answering questions about relations among relatives', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **6**, 479–491.

Potts, G. R. (1974), 'Storing and retrieving information about ordered relationships', *Journal of Experimental Psychology* **103**, 431–439.

Potts, G. R. (1976), 'Artificial logical relations and their relevance to semantic memory', *Journal of Experimental Psychology: Human Learning and Memory* **2**, 746–758.

Quayle, J. D. & Ball, L. J. (2000), 'Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning', *Quarterly Journal of Experimental Psychology* **53A**, 1202–1223.

Quinn, J. G. (1994), 'Towards a clarification of spatial processing', *Quarterly Journal of Experimental Psychology* **47A**, 465–480.

Quinn, S. & Markovits, H. (1998), 'Conditional reasoning, causality, and the structure of semantic memory: strength of association as a predictive factor for content effects', *Cognition* **68**, B93–B101.

Revlin, R., Leirer, V. O., Yopp, H. & Yopp, R. (1980), 'The belief bias effect in formal reasoning: The influence of memory on logic', *Memory & Cognition* **8**, 584–592.

Roberts, M. J. & Sykes, E. D. A. (2003), 'Belief bias and relational reasoning', *Quarterly Journal of Experimental Psychology* **56A**, 131–154.

Schaeken, W., Van der Henst, J.-B. & Schroyens, W. (in press), The mental models theory of relational reasoning: Premise relevance, conclusion phrasing and cognitive economy, *in* W. Schaeken, A. Vandierendonck, W. Schroyens & G. d'Ydewalle, eds, 'The Mental Models Theory of Reasoning: Extensions and Refinements', Lawrence Erlbaum Associates, Mahwah, NJ.

Siegler, R. S. & Shipley, C. (1995), Variation, selection, and cognitive change, *in* G. Halford & T. Simon, eds, 'Developing Cognitive Competence: New Approaches to Process Modeling', Erlbaum, Hillsdale, NJ, pp. 31–76.

Smyth, M. M. & Scholey, K. A. (1992), 'Determining spatial span: The role of movement time and articulation rate', *Quarterly Journal of Experimental Psychology* **45A**, 479–501.

Smyth, M. M. & Scholey, K. A. (1994), 'Interference in immediate spatial memory', *Memory & Cognition* **22**, 1–13.

Smyth, M. M. & Scholey, K. A. (1996), 'Serial order in spatial immediate memory', *Quarterly Journal of Experimental Psychology* **49A**, 159–177.

Sternberg, R. J. (1980), 'Representation and process in linear syllogistic reasoning', *Journal of Experimental Psychology: General* **109**, 119–159.

Sternberg, R. J. (1981), 'Reasoning with determinate and indeterminate linear syllogisms', *British Journal of Psychology* **72**, 407–420.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W. & Campbell, J. I. D. (2003), 'Syllogistic reasoning time: Disconfirmation disconfirmed', *Psychonomic Bulletin & Review* **10**, 184–189.

Toms, M., Morris, N. & Ward, D. (1993), 'Working memory and conditional reasoning', *Quarterly Journal of Experimental Psychology* **46A**, 679–699.

Van der Beken, H. & Vandierendonck, A. (2005), 'Why ancestors are treated differently than great-grandfathers in genealogical reasoning'. Manuscript submitted for publication.

Vandierendonck, A. & De Vooght, G. (1997), 'Working memory constraints on linear reasoning with spatial and temporal contents', *Quarterly Journal of Experimental Psychology* **50A**, 803–820.

Vandierendonck, A., Dierckx, V. & De Vooght, G. (2004), 'Mental model construction in linear reasoning: Evidence for the construction of initial annotated models', *Quarterly Journal of Experimental Psychology* **57A**, 1369–1391.

Wood, D. & Shotter, J. (1973), 'A preliminary study of distinctive features in problem solving', *Quarterly Journal of Experimental Psychology* **25**, 504–510.

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Mental Models in Learning Situations

Norbert M. Seel

Department of Educational Science, Universität Freiburg [1]

Abstract
Learning situations where phenomena are explained require the construction and successful manipulation of mental models. In these situations the models have the function to facilitate simplifcation and visualization of the modeled phenomena and the construction of analogies. This chapter reprots recent developments in research on model-centered learning with a focus on design-based modeling in the context of exploratory learning and guided discovery learning. Results of two large empirical studies on mental models in multimedia learning and discovery learning are reported.

1. Introduction

The idea of model-oriented learning has a long tradition in 20th century psychology and epistemology in which various roots can be distinguished. Bandura (1971) developed a paradigm for the field of social learning based on the imitation of a model's behavior. Craik (1943) introduced the idea of internal models to cognitive psychology with the notion of a working model. He argued that an individual who intends to give a rational explanation for something must develop practicable methods in order to generate adequate explanations from knowledge of the world and with limited capacities for

[1] E-mail: seel@uni-freiburg.de

information processing: Thus, in order to create situation-specific plausibility, the individual constructs a model that integrates the relevant semantic knowledge and meets the requirements of the situation to be mastered. This model "works" when it is within the realm of the subject's knowledge as well as the explanatory need with regard to the concrete learning situation to be mastered cognitively. A similar conception of internal models has been adapted by numerous psychologists who were concerned with the investigation of people's operations of complex technical or physical systems (see, for example Hacker 1977, Veldhuyzen & Stassen 1977).

Moreover, the conception of internal models also played a central role in information science in the 1950s and 1960s. Here we can find the idea that information exchange occurs by means of communication and information processing. Accordingly, learning was considered a complex procedure of information processing, and there were several authors, such as Steinbuch (1961), who considered the learning process as the procedure one uses to construct internal models of the environment. Such models are conceived as cognitive isomorphisms of structured domains or elements of the environment. The isomorphism is considered to be a threshold value which can be approached by the internal models of a subject but not reached.

As a result of the "cognitive revolution" in the 1960s (Bruner 1990), two decades later the theory of mental models became a very influential approach for both cognitive and educational psychology. The idea of mental models, which encompasses situated cognition as well as qualitative reasoning (Johnson-Laird 1983, Gentner & Stevens 1983, Greeno 1989), is based on two assumptions: (1) The person constructs a mental representation of reality, and (2) cognition and learning consist in the use of mental representations, in which individuals organize symbols of experience or thought in such a way that they effect a systematic representation of this experience or thought as a means of understanding it or of explaining it to others (Seel 1991). Learning occurs when people actively construct meaningful mental representations from information presented to them, such as coherent mental models that represent and communicate subjective experiences, ideas, thoughts, and feelings (cf. Mayer et al. 1999).

This chapter focuses on mental models constructed in learning situations to explain phenomena to be mastered cognitively. In the past, various instructional functions of models, such as envisioning and analogical reasoning, have been investigated in the fields of text and discourse processing (Rickheit & Habel 1999) and in the operation of complex systems of physics or economics (Markman 1998, Seel et al. 2000). Furthermore, there is a tradition of research on model building activities for specific disciplines, such as mathematics and science education. This research emphasizes design-based modeling in the context of guided discovery and exploratory learning (Lesh & Doerr 2000, Penner 2001). All of these movements can be subsumed under the broader field of model-centered learning.

2. What is model-centered learning?

From both a psychological and an epistemological point of view, a person constructs a model with a specific intention, i.e. in order to "map" the environment in a certain respect. In order to illustrate this one can refer to globes, which are models of the earth. Naturally, a particular globe is not a little earth. Rather, it is constructed and designed to give answers to questions concerning the locations of different places or distances between places. With regard to the chemical composition of the earth, a globe is not relevant. Other examples of modeling can be taken from the field of physics, such as Rutherford's atomic model or Newton's models of gravitation. These examples show that models are always representations of something: They represent natural or artificial objects, so-called originals, which can in their turn be models of something. Accordingly, talking about models implies, first of all, asking about the original to be modeled.

From the formal point of semantics, modeling can be defined as a homomorphism between relational systems. A relational system $\mathfrak{A} = [A, R_1A, \ldots, R_nA]$, i.e. the base domain or original, may be mapped on another relational system $\mathfrak{B} = [B, S_1B, \ldots, S_nB]$, i.e. the target domain, with the aim of explaining the target domain with the help of the base domain. In epistemology and cognitive psychology, this mapping is called an analogy and presupposes the construction of two internal models of these domains. This can be illustrated by an example provided by Holyoak & Thagard (1995, 33): "... [O]ur knowledge of water provides us with a kind of internal model of how it moves. Similarly, our knowledge of sound provides us with a kind of model of how sound is transmitted through the air. Each of these mental models links an internal representation to external reality. But when we consider the analogy between water waves and sound propagation, we are trying to build an isomorphism between two internal models. Implicitly, we are acting as if our model of water waves can be used to modify and improve our model of sound." The structural features of model building and the homomorphisms and isomorphisms involved with them have been described in more detail by Seel (1991). On the basis of these structural features, four functions of model building can be distinguished:

(1) Models 'aid in' the simplification of an investigation to particular and relevant phenomena in a closed domain.
(2) Models aid in the envisioning (or visualization) of a complex structure or system.
(3) Models aid in the construction of analogies which help to identify the structure of an unknown domain with the help of the structure of a known domain. In this way, a well-known explanation (e.g. Rutherford's atomic model) can be mapped onto a phenomenon to be ex-

plained (e.g. quantum mechanisms). Such models are called analogy models.

(4) Finally, models may aid in the simulation of the processes of a system. This occurs when an individual interacts with the objects involved in a situation in order to manipulate them mentally in such a way that the cognitive operations simulate specific transformations of these objects that may occur in real-life situations. These simulation models operate as thought experiments which produce qualitative inferences with respect to the situation to be mastered.

According to Stachowiak (1973), there are two main classes of mental models: perceptual models and thought models. Glaser et al. (1987) and Johnson-Laird (1983) refer to perceptual models as appearance or structural models that represent the external world in a static manner. This concept of appearance models corresponds to a great extent with the concept of models in information science (Weltner 1970). Thought models include qualitative process models as well as inductively derived artifacts that represent physical systems and their causal relationships in a dynamic manner. However, Norman (1983) has pointed out that we must distinguish between our conceptualization of a mental model and the actual mental model we think a person might have. To capture this idea, he separates the concept of "conceptual models" from that of "mental models." Accordingly, Kluwe & Haider (1990) distinguish between different kinds of models:

- Firstly, for a (complex) system S of the world there is a subjective internal or mental model of S, MM(S), which represents the knowledge a person has or can reconstruct with regard to S.
- Secondly, there is an "objective" model OM(S)—developed by scientists on the basis of their subjective mental models. We consider such models to be conceptual models, CM(S), and they represent the objective knowledge of a discipline. CM(S) can thus be conceived as the shared knowledge of a scientific community that results from the mental models of individual scientists.
- Thirdly, cognitive psychologists develop psychological models of the mental models of a system: PM[MM(S)]. These are the conceptual models referred to by Norman (1983).

Interestingly, Kluwe & Haider (1990) introduce a fourth kind of model that is especially important for instructional design: design and instructional models, DIM[CM(S)]. These models can be understood as instructionally designed conceptual models of a system S that are used for the construction of interfaces (learning tasks, manuals, and training) in order to guide the learners' construction of mental models. These "designed instructional models" are related to all other types of models. The relations can be illustrated as follows (Seel 2003):

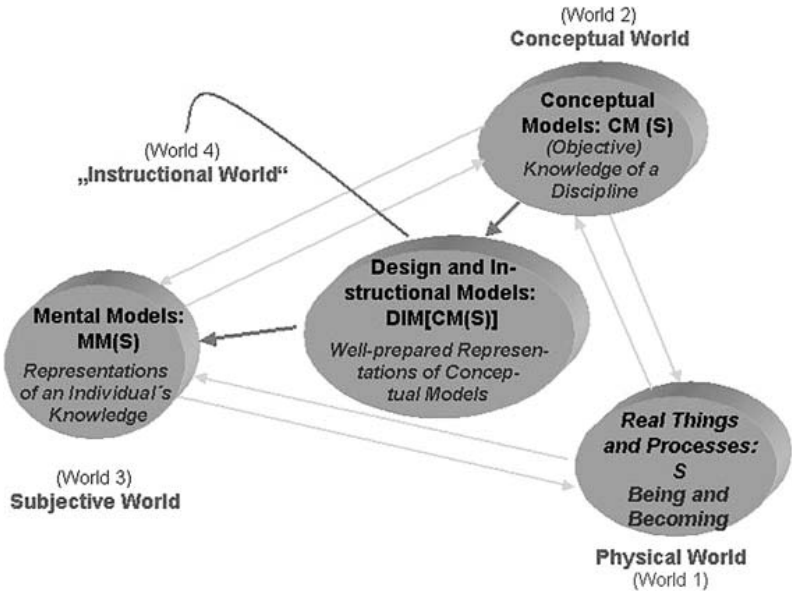


Fig. 1. The interplay of designed instructional models with other kinds of models

3. How can we influence model-centered learning through instruction?

The question of how we can influence model-centered learning through instruction has long been at the core of various educational approaches (see, for example, Karplus 1969), and in the field of research on mental models we can find a strong pedagogical impetus from the very beginning. According to Johnson-Laird (1989) and other authors, we can distinguish in principle between several sources for the construction of mental models: (1) the learner's ability to construct models in an inductive manner, either from a set of basic components of world knowledge or from analogous models that the learner already possesses; (2) everyday observations of the outside world in association with the adaptation of cultural models; and (3) other people's explanations. Among these sources, the third one seems to be especially relevant for education and instruction.

According to Carlson (1991), instruction can be designed to involve the learner in a process of inquiry in which facts are gathered from data sources, similarities and differences among facts noted, and concepts developed. In this process, the instructional program serves as a facilitator of learning for students who are working to develop their own answers to questions. On the other hand, instructional programs can present clearly defined concepts

followed by clear examples. A designed conceptual model may be presented ahead of the learning tasks in order to direct the learner's comprehension of the learning material. More generally, we can distinguish between different paradigms of model-centered instruction depending on whether they aim at (a) self-organized discovery and exploratory learning, (b) externally guided discovery learning, or (c) learning oriented toward the imitation of an expert's behavior or the adaptation of teachers' explanations.

Clearly, there might exist environments that can initiate a form of learning based on free exploration by invention, but in instructional contexts we regularly operate with well-prepared and designed learning environments that constrain the student's learning processes to various extents. Accordingly, at the beginning of research on model-centered instruction the focus was on the pedagogical idea as expressed by Mayer (1989, 47), which suggests that "students given model-instruction may be more likely to build mental models of the systems they are studying and to use these models to generate creative solutions to transfer problems." As a consequence, many studies on the learning-dependent progression of mental models have focused on the internalization of conceptual models provided to students in the course of instruction (Mayer 1989, Seel 1995, Seel et al. 2000). This research belongs to the third field listed above, learning oriented toward the imitation of an expert's behavior or the adaptation of teachers' explanations.

An alternative approach emphasizes the role of discovery learning for the construction of mental models (Penner 2001). According to this approach, the learner has to search continuously for information in a given learning environment in order to complete or stabilize an initial mental model which corresponds to an "a priori understanding" of the material to be learned. The goal of instruction is to create microworlds in which objects follow specific sets of rules. One example is a microworld in which balls fall in accordance with Newton's laws of motion (White 1993). Students explore this model by developing hypotheses and then varying input parameters to investigate how well their conjectures align with the model. In mathematics education the defining characteristic of this kind of discovery learning is that students explore conventional mathematical symbolizations in experientially real settings (Kaput 1994). More generally, Doerr (1996) states with regard to the various settings of discovery learning that students have to develop expressive models to explain phenomena using a variety of tools. According to Doerr, this model building begins with the students' informal understanding and progressively builds on it.

Self-guided learning occurs as a multi-step process of model building and revision (Penner 2001). Johnson-Laird (1983, 452) conceives this process as a "fleshing out" procedure which can be understood as a reductio ad absurdum that continuously examines whether a model can be replaced

with an alternative model or not (Seel 1991). Self-guided discovery learning requires some cognitive sophistication, i.e. learners must have previously achieved sufficient problem-solving and metacognitive skills to guide their learning process. Therefore, for beginning students it can be argued that self-organized discovery learning is closely associated with learning by trial-and-error but not by deep understanding. In addition, Briggs (1990) demonstrated in a case study that an instructional strategy aiming at discovery learning may dramatically increase the probability of stabilizing initial faulty mental models. Consequently, a substantial conceptual change does not take place, and relatively stable intermediate states of causal understanding often precede the conceptual mastery intended by instruction. In sum, self-organized learning aimed at the creation of mental models can indeed be rather precarious. It is a process that sometimes can even make an expert sweat. Thus, in order to be effective, learning environments aiming at model building activities must be designed carefully.

4. Designing effective environments for model-centered learning

Decades ago, Wertheimer (1959) pled for learning environments designed in such a way that learners can work on the solution of new problems effectively. In the 60s and 70s, several educational psychologists argued in a similar vein in accordance with Piaget's epistemology. For example, Bruner (1966) introduced the idea of guided discovery learning into the educational discussion, whereas Farnham-Diggory (1972) favored "free learning environments" and Stolurow (1973) developed his conception of transactional instruction, according to which learning environments should provide opportunities for reflective thinking. These different conceptions agree on the point that learning can be supported externally but not forced. Stolurow, for example, argues that if we want to improve exploratory learning and problem solving we need well-designed environments which provide the learners with optimal conditions for the development of initiatives and reduce external guidance to a minimum. From Stolurow's point of view, learning environments are not given a priori but rather must be developed and designed. Accordingly, he explicitly pleads for a program of instructional design as an evolving technology based on theoretical assumptions about psychological dispositions of the learner, learning activities, realistic learning results, and the potential effects of learning materials.

We can summarize these different lines of argument by stating that successful model-centered instruction presupposes that effective learning environments be designed in accordance with two different conceptions: First, there is a goal-oriented design of learning environments that has to be done

by instructional designers and aims at the internalization of well-designed conceptual models provided to the students. Second, there are instructional approaches which emphasize the self-organized construction and revision of models by students in the course of discovery learning. Gibbons (1998, 2002) integrates both lines of argumentation: "The events of instruction, which are the structures we design, serve human learning processes under the ultimate control of the individual. Instruction, therefore, does not cause learning but supports learning intentions the learner commits. [...] Some of these processes (such as the initial processing of visual or auditory information) are involuntary, but many of them (focusing attention, finding and selecting associations, etc.) are completely voluntary." (p. 3) In accordance with this precept, Gibbons formulates seven principles of model-centered instruction, which include (1) experience (i.e. learners should be given maximum opportunity to interact with one or more self-constructed models of systems for learning purposes), (2) problem solving, (3) denaturing, (4) sequence, (5) goal orientation, (6) resourcing, and (7) instructional augmentation. Gibbons intends for these seven principles to be considered as a fundamental basis for the instructional design of effective learning environments.

Several approaches of model-oriented teaching, such as the cognitive apprenticeship approach (see sec. 5 for details) or Gravemeijer's approach for mathematic education (Gravemeijer et al. 2000), correspond with Gibbons's principles. Gravemeijer argues that emergent models play a central role in individual students' learning and in the collective mathematical development of the classroom community. The notion of emergent models encompasses some aspects of the exploratory approach insofar as students are encouraged to develop their own models but do so in situations that are chosen by the teacher to support the realization of a proposed learning trajectory. Thus, it is possible for the designer to propose a developmental route for the classroom community in which students first model situations in an informal way (this is called a model of the situation) and then formulate their informal modeling activity mathematically (this produces a model for reasoning). Whereas Gravemeijer's approach can be situated between externally guided and discovery learning, another current movement of instructional research is closely related to the idea of model-based discovery learning. Bhatta & Goel (1997) have developed an interesting approach called "Integrated Design by Analogy and Learning (IDeAL)" as part of a theory of adaptive design. Similarly, Smith & Unger (1997) emphasize conceptual bootstrapping as a conception of analogy-based learning and problem solving. Both conceptions are based on the assumption that learners create their own designs through the retrieval of and adaptation to known designs.

An in-depth analysis of the various approaches of model-based discovery learning that aim to improve transfer between complex domains indicates that this kind of learning presupposes well-designed learning environments and materials. Bhatta & Goel (1997), for example, emphasize task-guided learning, which is dependent on designed learning tasks and the learner's domain-specific prior knowledge. Accordingly, the instructional design of learning tasks is at the core of IDeAL, which encourages students to construct device designs (in the fields of electricity, electronics, and heat exchangers) by having them carry out model-based and similarity-based learning which refers to retrievable knowledge about primitive functions within the known domain. Another approach of design-based modeling was developed by Erickson and Lehrer (1998). They distinguish between the design components planning, transforming, evaluation, and revision. Each of these components involves various model building activities. For example, planning includes defining the nature of the problem (asking questions) and managing the project (e.g. composition of the learning group and decision-making concerning tasks and roles), whereas transforming consists of information search, information extraction, organization, and so on.

5. Lessons learned from research

In the following paragraphs, several projects are introduced that focused on model-centered learning in various instructional settings. The first project, realized between 1994 and 2001,[2] was concerned with the internalization of pre-designed conceptual models provided to students in the course of multimedia learning. The focus of the second project[3] is on the use of mental models as devices for problem solving and discovery learning. The main characteristic of both projects is the strong orientation toward basic research on the learning-dependent progression of mental models initiated through instruction. More specifically, we designed the learning environments mainly in order to test theoretical hypotheses, our main interest being the systematic experimental variation of decisive model-building factors.

[2] We gratefully acknowlede financial support for this research from a generous grant by the Deutsche Forschungsgemeinschaft (German Research Association) with Grant-No. Se399/4. The research group consisted of Sabine Al-Diban, Susanne Held, Claudia Hess, Wolfram Lutterer, Christoph Nennstiel, Katharina Schenk, Ralph Siegel, and Susan Wilcek.

[3] This project started in March 2003. Again I gratefully acknowledge financial support for this research from a generous grant by the Deutsche Forschungsgemeinschaft (German Research Association) with Grant-No. Se399/8. The research group consists of Bettina Couné, Ulrike Hanke, Dirk Ifenthaler, Katharina Schenk, and Susanne Steiner.

5.1. Research on the assimilation of models

In the first comprehensive project, a series of replication studies was conducted to investigate externally guided model-based learning in a comprehensive multimedia learning environment designed in accordance with the principles of the cognitive apprenticeship approach (Collins et al. 1989). This approach proved to be a promising instructional strategy, providing students with pre-designed conceptual models to encourage them to imitate an expert's explanations. Moreover, this instructional approach prescribes in detail what the learner has to do in each sequence of learning in order to achieve particular objectives.

According to the cognitive apprenticeship approach, effective learning environments can be characterized by 18 features in four broad dimensions: content, methods, sequencing, and the sociology of teaching. Seel & Schenk (2003) separated a fifth dimension by emphasizing the important aspects of motivation and the corresponding need for a motivational design of learning environments. As cognitive apprenticeship is mainly concerned with macro-aspects of planning, we combined it in a further step with Jenkins' 1979 tetrahedral model, which we consider to be relevant for the micro-level of the design of learning tasks. The result of the combination of both approaches can be described as in Table 1 (cf. Seel & Schenk 2003). In modeling, an expert explains the conceptual model "economic circuit" and the cybernetic model "control loop." The students are instructed to adapt these conceptual models to accomplish the subsequent phases of learning. In coaching, the students are supervised and given guidance as they try to find solutions to a given task in an adaptive manner. The guidance given in coaching involves "result-oriented support" that was not difficult to realize, whereas in scaffolding "process-oriented support" was realized. Additionally, a special heuristics for problem solving was taught. This consisted of the decomposition of a complex problem into sub-problems and the construction of analogies between the sub-problems. Furthermore, two different instructional strategies for operating with analogies were realized: (1) subsumption of analogous learning tasks under the schema of a general problem solving structure, followed by its instantiation through a detailed and elaborated example; and (2) induction of a more general problem solving schema from analogous learning tasks through a comparison of different examples in order to extract structural similarities.

It turned out that realizing articulation and reflection within the multimedia program is a severe problem. Articulation is defined as the process of "thinking aloud" while working on a task, and reflection is defined as the comparison of the problem solving procedures applied by the learner and the expert. Collins et al. (1989) maintain that these methods contribute to

Table 1
Intersection between the cognitive apprenticeship approach and the Jenkin's tetrahedral model

Jenkins / Cognitive Apprenticeship	**Personal Variables**	**Learning Tasks Materials**	**Activities of learning**	**Results of learning, Criteria**
Contents	***Declarative knowledge Heuristic knowledge***	Curriculum of a subject matter Topic	**Acquisition of declarative and procedural knowledge**	Schemata Mental Models
Methods	*Control strategies Learning Styles*	***Modeling Coaching Scaffolding*** *Articulation Reflection Exploration*	Generative, procedural learning Metacognition	Rules, Principles, Proceduralization
Sequencing	***Knowldge Organization***	***Sequencing of learning steps***	***Increasing complexity and variety***	„Learning Hierarchies"
Motivation	*Intrinsic Motivation*	***Difficulty of tasks***	*Need for achievement*	***Interests Attitudes***
Sociology	*Cooperation Competition*	*Authenticity Contextuality*	*Culture of expert practice Team spirit*	Social behaviors and skills Attitudes

the development of reflective thinking and metacognitive control of learning. We realized both methods in the form of a "teach-back" procedure (Sasse 1991) in a social learning situation. This procedure is based on a "constructive interaction" between two communication partners who have similar domain-specific knowledge. One of them plays the role of a teacher who explains, for example, the states, functions, and transformations of a complex system to the other. The concrete task consisted in drawing causal diagrams (defined here as externalizations of mental models) (cf. Seel 1999).

In the final part of the apprenticeship instruction, called exploration, learners had to solve transfer tasks—one of them required a "near transfer" (i.e. the task remains in the same subject matter domain of economics) the other one required a "far transfer" from economics to ecology. All in all, the results of five evaluation studies with more than 400 subjects justify the statement that the cognitive apprenticeship approach is a sound framework for the instructional design of environments for constructivist learning. So far these results correspond with observations and empirical results of other studies, such as those of Casey (1996) and Chee (1995).

However, as in Casey's study it was difficult to realize the methods of articulation and reflection in a multimedia learning environment. Basically, the same holds true with respect to the realization of scaffolding. Nevertheless, the effectiveness of the multimedia program for learning is empirically well substantiated with regard to apprenticeship methods which aim at explanatory descriptions, as in expository teaching. The sequence of the methods "modeling – coaching – scaffolding – exploration" significantly improved the learning process, as did the accomplishment of the complex transfer tasks in exploration.

Apart from this, the overall results with regard to the effectiveness of the apprenticeship methods suggest a more detailed analysis of the learning and transfer performances in order to separate the methods' effectiveness (for more details: Seel & Schenk 2003). The weak spot of the instruction was scaffolding, the fact that none of our efforts to enable the learners to develop a promising problem solving strategy (e.g. decomposing a complex problem into sub-problems and solving them by analogy) were effective. The significant decrease in performance between coaching and scaffolding that was observable in all replication studies indicates that learners could not progress from content-oriented to process-oriented learning in the sense of an increasingly self-regulated accomplishment of analogous tasks (cf. Alexander et al. 1987, Newby et al. 1995). An explanation for this may be found in the fact that the subjects of our studies were constrained by the instructional program and did not receive additional advice by a teacher as suggested by Palincsar (1986), who considers dialogue to be a solid basis for effective scaffolding. The multimedia instruction was not capable of adapting the learning tasks to the individual learner. For example, it can not adapt the difficulty of a learning task to the learners' abilities to compensate for a learner's missing knowledge. Furthermore, the multimedia instruction did not make appropriate "cognitive tools" available to support learners in accomplishing the learning tasks. Actually, learners were provided with an easier learning task which they could solve and then the difficulty of tasks increased until the learners were no longer able to solve them on their own. Hmelo & Guzdial (1996) view this organization of tasks as an example of "black-box scaffolding," which may improve the performance in the case of "closed" domains but is ineffective in making the intended scaffold transparent for problem solving. Obviously, our data confirm this argumentation. As an alternative, Hmelo and Guzdial consider redesigning learning tasks to support task performance with the help of a "supplantation" (as defined by Salomon 1979) of the cognitive operations involved in the task solutions. Moreover, task performance can be supported by the application of cognitive tools that give advice to learners on how to represent and manipulate a problem (for example, with the help of graphic diagrams). These forms of scaffolding are taken by Hmelo and Guzdial to be exam-

ples of "glass-box scaffolding," as their aim is to help learners on problems they cannot master on their own. In accordance with this argumentation, we asked the learners to draw causal diagrams of their mental models at different time points in the course of learning. Actually, a central goal of the project was the assessment of the learning-dependent progression of mental models. Therefore, the learning outcomes could also be evaluated by means of causal diagrams, which can be considered as a combination of cognitive modeling and a particular structure-spreading technique similar to concept mapping (for more details, see Seel 1999). Whereas the learners could improve their domain-specific knowledge only slightly in the various replication studies, we were able to observe significant changes not only in the accomplishment of the various learning tasks but also in the causal diagrams drawn at various points of measurement. This specific kind of knowledge diagnosis confirmed central assumptions of the theory of mental models (cf. Seel et al. 2000). Actually, the results of the various replications studies indicate that causal diagrams can be considered suitable methods to assess mental models as knowledge constructions of higher order that develop in dependence on current learning experiences. The quality of the causal diagrams (i.e. both their associative strength and their complexity) improved in the course of instruction. We were thus able to interpret the substantial changes in causal diagrams we observed as "evidence" for the general effectiveness of the instructional intervention. Obviously, the effective design of successful learning environments presupposes the provision of cognitive tools which facilitate and support individual model building and revision with the goal of problem solving. With regard to the theory of mental models, the results of these investigations support the assumption that mental models—as measured with the help of causal diagrams—are situation-dependent constructions (cf. Seel 2001). To remain within the same context of contents, the learners did construct—at different times—causal diagrams as cognitive artifacts which correlated only minimally with each other. Obviously, it was more parsimonious and cognitively less exhaustive to construct a new causal diagram at each time of measurement than to remember the previously constructed solutions. Even in cases where the learners modified a previously created causal diagram, the observable changes were so substantial that the result was a new causal diagram. We interpret these relatively consistent results as "evidence" for the specific function of mental models in aiding situated cognition. From the perspective of instructional research, the results of our investigations contradict to a great extent the widely accepted assumption that students adapt externally provided models and apply them to solve tasks. Actually, in the various replication studies the learners' causal diagrams offered only minor similarities with the conceptual models provided in instruction. Although contingency coefficients indicated that the learners' causal diagrams were

not fully independent of the conceptual models, the correlations were not significant. Basically, we can agree with the verdict of Mayer (1989, 47) that "students given model-instruction may be more likely to build mental models of the systems they are studying and to use these models to generate creative solutions to transfer problems," but at the same time it is clear that the students do not adapt the conceptual model provided in instruction one-to-one. Rather, in the course of learning with the instructional program they acquired domain-specific knowledge which they used for the construction of independent causal models. This corresponds with the postulate of constructivist learning that the learning environment provided is an important informational resource that can be used in a strategic manner to extract the information needed to create subjective plausibility and solve learning tasks.

5.2. Model-based problem solving and discovery Learning

Parallel to the instructional research inspired by the mental model approach, we can find—especially in the fields of mathematics and physics education—various approaches of model-centered instruction. Stewart et al. (1992) have circumscribed the central idea of these instructional approaches, stressing that "a science education should do more than instruct students with respect to the conclusions reached by scientists; it should also encourage students to develop insights about science as an intellectual activity" (p. 318). Accordingly, advocates of this approach argue that "given that we wish to involve students in the practices of scientists, we focus primarily on model building" (Penner et al. 1998, 430). Indeed, some of the most important goals of instruction in science are to help students develop powerful models to make sense of their experiences involving light, gravity, electricity, and magnetism. It is also obvious that young students invent models of their own and that changing their ways of thinking must involve challenging and testing these models (Penner 2001). The model building approach provides a significant challenge for the understanding of how to nurture, accommodate, and respond to the partial and incomplete models that students are likely to build with regard to phenomena of physics. We find a similar argumentation with regard to the learning of mathematics: "The primary role of algebra at the school level is to develop confidence and facility in using variables and functions to model numerical patterns and quantitative relationships." (National Council of Teachers of Mathematics, 1994) Accordingly, Lesh & Doerr (2000) and other authors talk about models that students should develop in attempts to produce mathematical descriptions or explanations of systems in the physical world. These

authors argue that helping students to develop powerful models should be among the most important goals of science and mathematics instruction.

At the moment my research group is involved in a comprehensive project in accordance with these conceptions on model-based discovery learning. The focus of this project is again on the assessment of the learning-dependent progression of mental models in the course of complex problem solving. This was realized in the context of a computer-based multimedia learning environment designed in accordance with the approach of model-centered learning and instruction (Seel 2003). The learning environment is modular in structure and can be divided into declarative and heuristic modules.

The declarative modules contain all information needed to solve the phenomenon in question. The heuristic modules primarily support the model building process and analogical reasoning. However, the essential heuristic module is the so-called Model-Building-Kit (MoBuKi), which provides students with information about models, model building, and analogical reasoning as well as with examples of analogies applied successfully on a meta-level. In this respect, the MoBuKi offers a heuristics for problem solving which can be transferred to various contents. In addition, four supplementary modules complete the learning environment. A curriculum module contains scientific information on the prevailing content. Here the learners can navigate through different topics. However, there are no models available within this module, and thus learners have to construct their own models using the information provided. The module "wissen.de" includes various text documents, audio recordings, and pictures to complement the information in the curriculum module. Another module is the presentation of the problem and learning task, where the learners are requested to solve a complex problem. The task the students are provided with is to construct two models—one model which explains the problem (explanation model) and a second model with relations and functions similar to the explanation model, which we call an analogy model. The toolbox "MS PowerPoint" is the module in the multimedia learning environment that allows students to externalize their mental models on the problem they are trying to solve. As the first step in the measurement of the learning-dependent progression of mental models, we focused on changes of semantic sensitivity in the student models. Accordingly, we measured the students' models at predefined stages of their learning process. To date, we have conducted two comparable studies with different disciplines (ecology and geophysics). In both studies we experimentally varied two factors of model-centered discovery learning: (1) individual vs. collaborative learning and (2) self-guided vs. scaffolding-based learning.

In a first study, 52 secondary school students (9th grade) took part in the experiment. The discipline of this study was geology. Due to the ex-

perimental variation, 26 students took part as individual learners whereas the remaining students worked as collaborative learners. We selected the stored student models of the individual learners in order to measure their learning-dependent progression. To indicate whether there was a change in the learners' models, we asked external "model raters" to determine whether there were similarities or differences in the structures of the models produced by the learners. More specifically, their task was to construct two models, one which explained the phenomenon in question (i.e. the so-called explanation model) and one with similar relations and functions (the so-called analogy model). In consequence, the "model raters" had to evaluate a total of 416 student models. In total, more than 50 "model raters" (separated into two independent groups) evaluated the learners' explanation and analogy models. In order to control the reliability of the ratings, we had the subjects evaluate the same set of models on two separate days four days apart to avoid a strong recognition effect. The explanation and analogy models of the students were put into chronological order and the "model raters" had to compare—by means of a questionnaire—the similarities or dissimilarities from different stages of the students' learning process. The first comparison consisted of the learners' preconception model (the "a priori" model, constructed before they worked with the multimedia learning environment) and the model constructed after the first day working with the multimedia learning environment. Comparisons 2 to 6 consisted of the models constructed during the subsequent work with the learning environment (32; 42; 52; 62; 72; 81). The learners were allowed to continue the learning process with the model from the preceding day. The last comparison consisted of the last model constructed while working in the learning environment and the so-called take-home model (8th), which the learners constructed on the last day without using the learning environment or the preceding models.

Results The coefficient of internal consistency calculated for the first evaluation group, MRem1, using Cronbach's coefficient alpha was .84 (n = 5642). For the second evaluation group, MRam1, the coefficient of internal consistency calculated using Cronbach's coefficient alpha was .86 (n = 3822). These findings provide evidence that our instrument can measure similarities or differences between models reliably. As expected, the probability of change between the preconception model (2pc) and the first learning day model (32) is very high ($p2pc_{32} = .99$). Between the last learning day model (81) and the take-home model (8th) there is also a high probability of change ($p81_{8th} = .80$). Between the first and the last learning day the probability of change decreased at an average of 16.9% per day (cf. table 2).

Table 2
Average probability of change in explanation models (n= 26 students)

measuring point (mp)	∅ probability of change
mp 2pc - mp 32	0.99547511
mp 32 - mp 42	0.77375566
mp 42 - mp 52	0.57013575
mp 52 - mp 62	0.30542986
mp 62 - mp 72	0.27149321
mp 72 - mp 81	0.15158371
mp 81 - mp 8th	0.80090498

The probabilities of change were partitioned into two groups on the basis of the experimental variation (scaffolding-based vs. self-guided learning) and entered into a one-way ANOVA. The analysis revealed a significant effect for the comparison of the models between measuring point 62 and 72, $F = 11.45$, $p < 0.05$ (see Figure 2). Evidently, this result can be explained by the instructional intervention given in the scaffolding-based learning group immediately before measurement 62. Beyond this effect, a one-way ANOVA showed no further significant differences between the two learning groups (scaffolding based vs. self-guided) on the preceding or following measuring points. The results for the analogy models revealed a slightly different picture (cf. figure 3). Unlike the results for the explanation models, the probability of change between the first and the last learning day for the analogy models did not show a continuous decrease (cf. table 3). Interestingly, the students change their mental models with a higher probability from measuring point 42 to 52 ($p42_{52} = .65$) than on the preceding or following measuring points. Again, the probability of change from the last learning day model (81) to the take-home model (8th) is very high ($p81_{8th} = .86$).

Interestingly, we were able to replicate these results in a second study attended by 79 secondary school students. As in the first study, the student's task in the second study was to construct two models (for explanation and analogy). Therefore, in study 2, the "model raters" had to compare a total of 462 student models. Indeed, for the explanation models we found a similar pattern of probabilities of change between the different measurements (e.g., $p2pc_{31} = .87$, and $p51_{5th} = .61$). Moreover, we found a continuous decrease of the probability of change between the five models constructed by the students while working with the multimedia learning environment. A comparison between the experimental treatments resulted in a significant

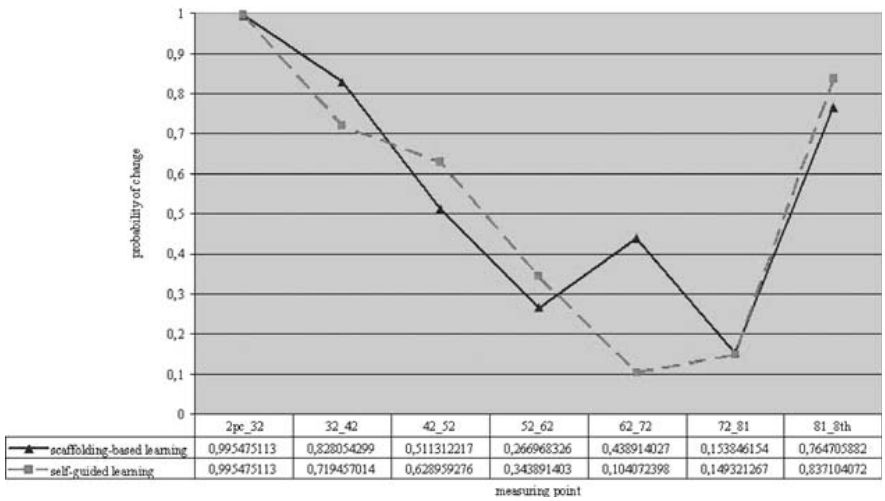


Fig. 2. Measurement of change for the explanation models of study 1

effect between the scaffolding-based learners and the self-guided learners between measuring points 32 and 41 (ANOVA, F = 4.62, p < .05). At this point, the probability of change in the scaffolding-based group ($p32_{41(sb)}$ = .552) turned out to be significantly higher than in the self-organized group ($p32_{41(so)}$ = .271). Again, this can be explained as a result of the instructional intervention before the measurement. Beyond this result, a one-way ANOVA showed no further significant differences between the two learning groups (scaffolding-based vs. self-guided).

Unlike the preceding results of the first study, the analogy models of the second study revealed a different picture upon analysis (see Figure 4). The results again showed a significant difference between measuring points 32 and 41 (ANOVA, F = 4.87, p < .05), where the probability of change in the scaffolding-based group ($p32_{41(sb)}$ = .591) was higher than that in the self-guided group ($p32_{41(so)}$ = .288) due to the instructional intervention.

Interpretation We can interpret the results of both studies as indicative for effects of a semantic sensitivity with regard to specific cues within a learning environment. The concept of semantic sensitivity was introduced by Anzai & Yokoyama (1984), who argued that individuals working on a learning task immediately encode the information on a task onto a mental model in order to generate a basic understanding of the situational demands. The concept is based on the capability of individuals to focus on cues in the

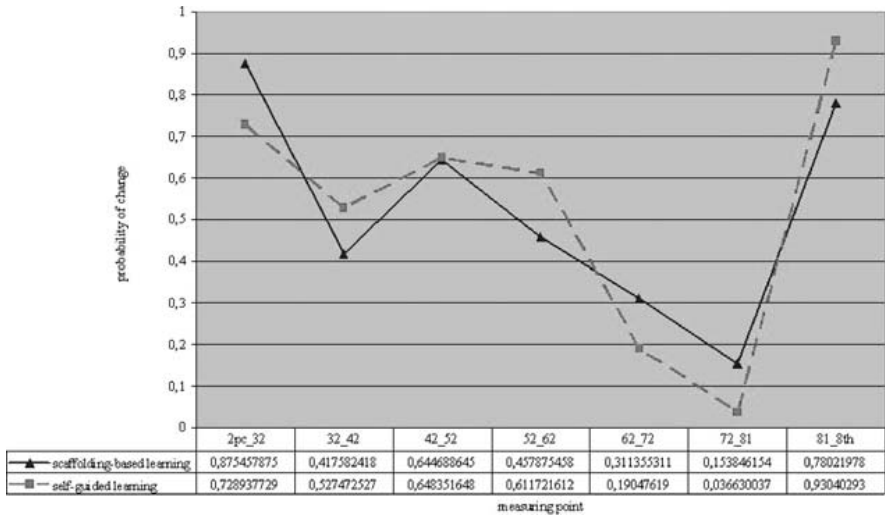


Fig. 3. Measurement of change for the analogy models of study 1

learning environment relevant for the model and to use them to construct a new mental model of the task that results in a more correct or better solution than the preceding model. This argumentation corresponds with the concept of cognitive reconstruction (cf. Dole & Sinatra 1998) as well as with earlier studies on the construction and revision of models in learning situations (cf. Seel 1995, Seel & Dinter 1995).

References

Alexander, P. A., White, C. S., Haensly, P. A. & Grimmins-Jeanes, M. (1987), 'Training in analogical reasoning', *American Educational Research Journal* **24**, 387–404.

Anzai, Y. & Yokoyama, T. (1984), 'Internal models in physics problem solving', *Cognition and Instruction* **1**, 397–45.

Bandura, A. (1971), *Social Learning Theory*, General Learning Press, New York.

Bhatta, S. & Goel, A. (1997), 'Learning generic mechanisms for innovative strategies in adaptive design', *The Journal of the Learning Sciences* **6**(4), 367–396.

Briggs, P. (1990), The role of the user model in learning as an internally and externally directed activity, *in* D. Ackermann & M. Tauber, eds, 'Mental Models and Human-Computer Interaction 1', Elsevier, Amsterdam, pp. 195–208.

Bruner, J. (1966), *Toward a Theory of Instruction*, Harvard University Press,

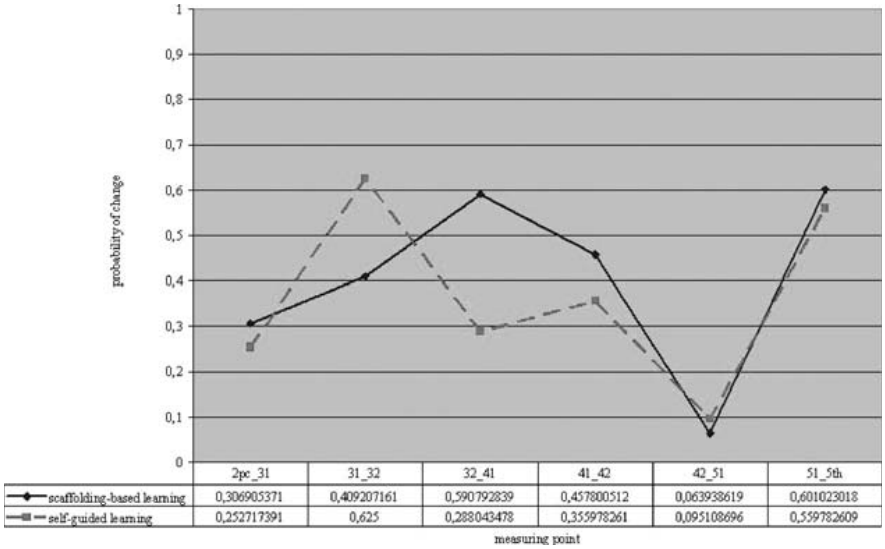


Fig. 4. Measurement of change for the analogy models of study 2

Cambridge, MA.

Bruner, J. (1990), *Acts of Meaning*, Harvard University Press, Cambridge, MA.

Carlson, H. (1991), 'Learning style and program design in interactive multimedia', *Educational Technology Research and Development* **39**(3), 41–48.

Casey, C. (1996), 'Incorporating cognitive apprenticeship in multimedia', *Educational Technology Research and Development* **44**(1), 71–84.

Chee, Y. (1995), 'Cognitive apprenticeship and its application to the teaching of smalltalk in a multimedia interactive learning environment', *Instructional Science* **23**, 133–161.

Collins, A., Brown, J. & Newman, S. (1989), Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics, *in* L. Resnick, ed., 'Knowing, Learning, and Instruction', Erlbaum, Hillsdale, NJ, pp. 453–494.

Craik, K. (1943), *The Nature of Explanation*, Cambridge University Press, Cambridge.

Doerr, H. (1996), 'Integrating the study of trigonometry, vectors, and force through modeling', *School Science and Mathematics* **96**, 407–418.

Dole, J. & Sinatra, G. (1998), 'Reconceptualizing change in the cognitive construction of knowledge', *Educational Psychologist* **33**(2/3), 109–128.

Farnham-Diggory, S. (1972), *Cognitive Processes in Education: A Psychological Preparation for Teaching and Curriculum Development*, Harper & Row, New York.

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

Gibbons, A. (1998, 2002), Model-centered instruction. Paper presented at the Annual Meeting of the American Education Research Association, San Diego,

CA, April 1998.

Glaser, R., Lesgold, A. & Lajoie, S. (1987), Toward a cognitive theory for the measurement of achievement, *in* R. Ronning, J. Glover, J. Conoley & J. Witt, eds, 'The Influence of Cognitive Psychology on Testing and Measurement', Lawrence Erlbaum, Hillsdale, NJ, pp. 41–85.

Gravemeijer, K., Cobb, P., Bowers, J. & Whitenack, J. (2000), Symbolizing, modeling, and instructional design, *in* P. Cobb, E. Yackel & K. McClain, eds, 'Symbolizing and Communicating in Mathematics Classrooms: Perspectives on Discourse, Tools, and Instructional Design', Erlbaum, Mahwah, NJ, pp. 225–273.

Greeno, J. (1989), Situations, mental models, and generative knowledge, *in* D. Klahr & K. Kotovsky, eds, 'Complex Information Processing', Erlbaum, Hillsdale, NJ, pp. 285–318.

Hacker, W. (1977), Bedeutung und Analyse des Gedächtnisses für die Arbeits- und Ingenieurpsychologie — zu Gedächtnisanforderungen in der psychischen Regulation von Handlungen, *in* F. Klix & H. Sydow, eds, 'Zur Psychologie des Gedächtnisses', Huber, Bern, pp. 150–174.

Hmelo, C. E. & Guzdial, M. (1996), Of black and glass boxes: Scaffolding for doing and learning, *in* 'Proceedings of the Second International Conference on the Learning Sciences', Association for the Advancement of Computers in Education, Charlottesville, VA, pp. 128–133.

Holyoak, K. & Thagard, P. (1995), *Mental Leaps: Analogy in Creative Thought*, The MIT Press, Cambridge, MA.

Jenkins, J. (1979), Four points to remember: A tetrahedral model of memory experiments, *in* I. L. Cermak & F. Craik, eds, 'Levels of Processing in Human Memory', Lawrence Erlbaum, Hillsdale, NJ, pp. 429–446.

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

Johnson-Laird, P. (1989), Mental models, *in* M. Posner, ed., 'Foundations of Cognitive Science', The MIT Press, Cambridge, MA, pp. 469–499.

Kaput, J. (1994), The representational roles of technology in connecting mathematics with authentic experience, *in* R. Biehler, R. Scholz, R. Sträßer & B. Winkelmann, eds, 'Didactics of Mathematics as a Scientific Discipline', Kluwer, Dordrecht, pp. 379–397.

Karplus, R. (1969), *Introductory Physics: A Model Approach*, Benjamins, New York.

Kluwe, R. & Haider, H. (1990), 'Modelle zur internen Repräsentation komplexer technischer Systeme', *Sprache & Kognition* **9**(4), 173–192.

Lesh, R. & Doerr, H. (2000), Symbolizing, communicating, and mathematizing: Key components of models and modeling, *in* E. Y. P. Cobb & K. McClain, eds, 'Symbolizing and Communicating in Mathematics Classrooms: Perspectives on Discourse, Tools, and Instructional Design', Erlbaum, Mahwah, NJ, pp. 361–383.

Markman, A. (1998), *Knowledge Representation*, Erlbaum, Mahwah, NJ.

Mayer, R. (1989), 'Models for understanding', *Review of Educational Research* **59**(1), 43–64.

Mayer, R., Moreno, R., Boire, M. & Vagge, S. (1999), 'Maximizing construc-

tivist learning from multimedia communication by minimizing cognitive load', *Journal of Educational Psychology* **91**(4), 638–643.

Newby, T., Ertmer, P. & Stepich, D. (1995), 'Instructional analogies and the learning of concepts', *Educational Technology Research and Development* **43**(1), 5–18.

Norman, D. (1983), Some observations on mental models, *in* D. Gentner & A. L. Stevens, eds, 'Mental Models', Erlbaum, Hillsdale, NJ, pp. 7–14.

Palincsar, A. (1986), 'The role of dialogue in providing scaffolded instruction', *Educational Psychologist* **21**(1/2), 73–98.

Penner, D. (2001), 'Cognition, computers, and synthetic science: Building knowledge and meaning through modeling', *Review of Research in Education* **25**, 1–35.

Penner, D., Lehrer, R. & Schauble, L. (1998), 'From physical models to biomechanics: A design-based modeling approach', *The Journal of the Learning Sciences* **7**(3/4), 429–449.

Rickheit, G. & Habel, C., eds (1999), *Mental Models in Discourse Processing and Reasoning*, Elsevier, Amsterdam.

Salomon, G. (1979), *Interaction of Media, Cognition and Learning*, Jossey Bass, San Francisco.

Sasse, M. (1991), How to t(r)ap users' mental models, *in* M. Tauber & D. Ackermann, eds, 'Mental Models and Human-computer Interaction 2', North-Holland, Amsterdam, pp. 59–79.

Seel, N. (1991), *Weltwissen und mentale Modelle*, Hogrefe, Göttingen.

Seel, N. (1995), 'Mental models, knowledge transfer, and teaching strategies', *Journal of Structural Learning* **12**(3), 197–213.

Seel, N. (1999), 'Educational diagnosis of mental models: Assessment problems and technology-based solutions', *Journal of Structural Learning and Intelligent Systems* **14**(2), 153–185.

Seel, N. (2001), 'Epistemology, situated cognition, and mental models: "Like a bridge over troubled water"', *Instructional Science* **29**(4-5), 403–428.

Seel, N. (2003), 'Model-centered learning and instruction', *Technology, Instruction, Cognition, and Learning* **1**(1), 59–85.

Seel, N., Al-Diban, S. & Blumschein, P. (2000), Mental models and instructional planning, *in* M. Spector & T. M. Anderson, eds, 'Integrated and Holistic Perspectives on Learning, Instruction and Technology: Understanding Complexity', Kluwer, Dordrecht (NL), pp. 129–158.

Seel, N. & Dinter, F. (1995), 'Instruction and mental model progression: Learner-dependent effects of teaching strategies on knowledge acquisition and analogical transfer', *Educational Research and Evaluation* **1**(1), 4–35.

Seel, N. & Schenk, K. (2003), 'Multimedia environments as cognitive tools for enhancing model-based learning and problem solving: An evaluation report', *Evaluation and Program Planning* **26**(2), 215–224.

Smith, C. & Unger, C. (1997), 'What's in dots-per-box? Conceptual bootstrapping with stripped-down visual analogies', *The Journal of the Learning Sciences* **6**(2), 143–181.

Stachowiak, H. (1973), *Allgemeine Modelltheorie*, Springer, Wien.

Steinbuch, K. (1961), *Automat und Mensch. Über menschliche und maschinelle Intelligenz*, Heidelberg.

Stewart, J., Hafner, R., Johnson, S. & Finkel, E. (1992), 'Science as model building: Computers and high-school genetics', *Educational Psychologist* **27**(3), 317–336.

Stolurow, L. (1973), Lernumwelten oder Gelegenheiten zum Nachdenken, *in* W. Edelstein & D. Hopf, eds, 'Bedingungen des Bildungsprozesses. Psychologische und pädagogische Forschungen zum Lehren und Lernen in der Schule', Klett, Stuttgart, pp. 351–398.

Veldhuyzen, W. & Stassen, H. (1977), 'The internal model concept: An application to modelling human control of large ships', *Human Factors* **19**, 367–380.

Weltner, K. (1970), *Informationstheorie und Erziehungswissenschaft*, Schnelle, Quickborn.

Wertheimer, M., ed. (1959), *Productive Thinking*, Harper & Row, New York.

White, B. (1993), 'ThinkerTools: Causal models, conceptual change, and science education', *Cognition and Instruction* **10**(1), 1–100.

This Page is Intentionally Left Blank

Part II

Cognitive Neuroscience

Introduction: Cognitive Neuroscience

In the last decade, a few research groups have attended to the question of how thinking with mental models is biologically realized in the human brain. Such neuro-cognitive investigations are performed by using modern brain imaging methods that enable researchers to monitor the brain at work. Functional magnetic resonance imaging (fMRI)—which is used in the studies presented here—takes advantage of the fact that cognitive processes lead to a local increase of oxygen in the activated cerebral tissue. Physically, the fMRI technique relies on the fact that deoxyhemoglobin is paramagnetic relative to oxyhemoglobin and the surrounding brain tissue, and that a local increase in oxygen delivery is correlated with brain activation. The principle of fMRI experiments is to measure brain activation repeatedly in short intervals and to explore differences among the activation patterns measured. Typically, the baseline activity is measured when the volunteer is at rest and other measurements are taken when he or she performs certain cognitive tasks. In the simplest experimental design, the activity in the baseline condition is then subtracted from the activity measured during the performance of the cognitive tasks. The resulting data can be statistically analyzed. Areas in which statistically significant differences were measured are presumed to have been activated by the cognitive task. In more sophisticated experiments, combinations of experimental conditions are compared to other combined conditions. To illustrate the results, the patterns of activation are usually transferred into so-called fMRI images, in which the most visible regions correspond to the areas activated by the cognitive task. The imaging technique brings the acquisition time for one image down to milliseconds, so that a whole brain can be scanned within a few seconds.

There are two issues to research on mental models that are difficult to answer solely on the basis of behavioral data. The first concerns the difference between reasoning with determinate and indeterminate problems. If the premises of an argument describe a situation unambiguously, exactly

one mental model can be constructed. In this sense, determinate problems determine one model. On the other hand, if a set of premises allows several interpretations more than one model conforms to the set. Hence, these cases are called "indeterminate problems." The other central issue is the relation between mental models and images (for a philosophical discussion of this issue, see part IV). In the following part, two papers are presented that approach these classical issues with neuro-cognitive methods.

Mental model theory assumes that mental models are processed in areas of the brain that are related to visuo-spatial information processing. Therefore, reasoning with multiple mental models should lead to an increase of blood flow in these areas as compared to reasoning with single mental modles. However, another thesis states that some sort of linguistic representation of the premises is kept in memory while constructing different possible models. In this case, multiple mental models would lead to an additional acitvation in linguistic areas. In addition to the behavioral studies conducted so far, **Waechter and Goel** explore the differences in brain activation during reasoning with determinate and indeterminate problems. The difference between determinate and indeterminate problems consists in the activation in both the left superior parietal cortex and the left frontal and temporal cortex. Since the first region can be related to spatial processing whereas the second region is involved in language processing, these data clearly support the second hypothesis. They conclude that any mental models is augmented with linguistic representations as soon as the reasoning problem becomes difficult.

The second issue, namely the relation between mental models and mental images, is investigated by **Knauff**. First, he shows that visual areas are only activated when the reasoning problem contains relations that are easy to visualize. In other reasoning problems, however, mental images do not play any role. Furthermore, his fMRI-data support the hypotheses that reasoning with mental models and visual images takes place in three steps: 1) visual image construction, 2) image to model transformation, and 3) mental model processing. In the course of reasoning, the activation in the brain moves from occipito-temporal regions, which are known to be involved in visual imagery, to anterior prefrontal regions, which play a major role in relational integration, and further to the posterior parietal cortex, which is the place where abstract spatial information is processed. Only the abstract spatial representation is used in the reasoning process.

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Resolving Valid Multiple Model Inferences Activates a Left Hemisphere Network

Randall L. Waechter and Vinod Goel [1]

Department of Psychology, York University, Toronto, Ontario [2]

Abstract

Resolving multiple model syllogisms is more difficult than resolving single model syllogisms. Mental model theory predicts that visuospatial processing is critical for resolving syllogisms, and that demands on visuospatial processing systems will increase as reasoning problems become more difficult. An alternative account, the mixed-model approach, postulates that linguistic representations may augment visuospatial representations in multiple model problems. To test these competing hypotheses, we reorganized published archival fMRI data into single and multiple model problems, and reanalyzed it along this dimension. The critical comparison of multiple model versus single model problems revealed activation in both the left superior parietal spatial system and left frontal and temporal language areas, indicating that as reasoning problems become more difficult, reasoners augment any visuospatial model that they may have constructed with linguistic representations. This result is consistent with the mixed-model approach.

[1] V.G. is supported by a McDonnell-Pew Program in Cognitive Neuroscience Award, NSERC and CIHR grants, and a Premier's Research Excellence Award.
The authors would like to thank Dr. Oshin Vartanian for his invaluable insight and assistance with the writing of this manuscript.
[2] E-mail: vgoel@yorku.ca

1. Introduction

Reasoning is the cognitive activity of drawing inferences from given information. Arguments are considered valid only if the information contained in the premises provides absolute grounds for accepting the conclusion. One influential theory of logical reasoning, mental model theory, claims that determining the validity of logical arguments requires "the understanding of discourse (that) leads to a model of the relevant situation akin to one created by perceiving or imagining events instead of merely being told about them" (Johnson-Laird 1995, 999). Consider the following categorical syllogism:

A. All California snails are amphibians.
No amphibians can sing.
Therefore, no California snails can sing.

In the above example, individuals might mentally construct the following representation of the relationship between the premises and conclusion (Fig. 1): Mental model theory postulates that the reasoner determines whether

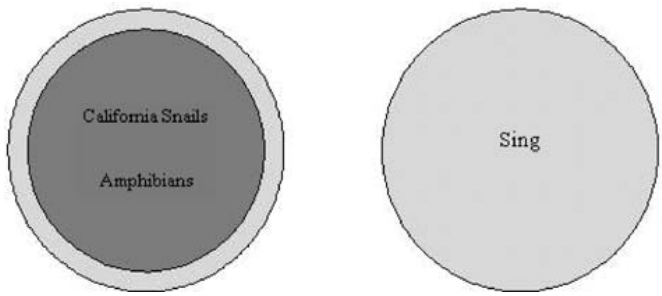


Fig. 1. Venn circle diagram of a single model syllogism

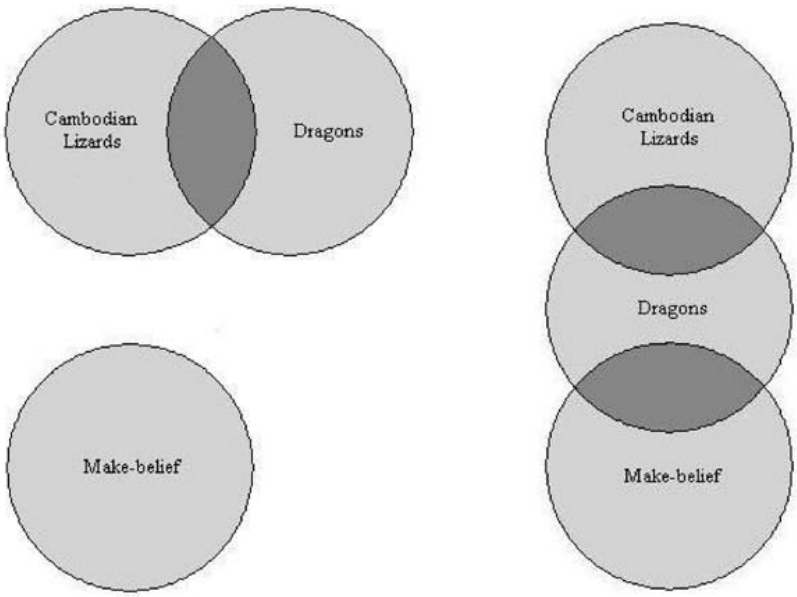
the syllogism is valid or not by examination of such a spatially organized model. Specifically, the validity of the argument is tested by searching for alternative permutations of the first two premises that refute the conclusion. In the above example, the reasoner attempts to visualize the premises "All California Snails are amphibians" and "No amphibians can sing" in some other way than that pictured in Figure 1. In this case, the reasoner determines that the "California snails" and "Amphibians" circles *must completely* overlap to indicate that *all* of the California snails are amphibians, while the "Amphibians" and "Sing" circles *must be completely separate* to indicate that *no* amphibians can sing. As there is only one permutation of the premises in this example, and it is consistent with the conclusion,

the argument must be valid. In fact, 90% of people given this particular syllogism draw the correct conclusion.

Now consider the following syllogism:

B. No Cambodian lizards are make-belief.
Some Cambodian lizards are dragons.
Therefore, some dragons are not make-belief.

Evaluation of this argument may result in the mental construction of a model like in Figure 2a. An important distinction can be made between



a. Reasoning Model Permutation # 1 b. Reasoning Model Permutation # 2

Fig. 2. Venn circle diagram of a valid multiple model syllogism

syllogism A and B. The premises in syllogism A can only be arranged in one way, making it a 'single model' syllogism. The premises in the second argument can be arranged in more than one way, making it a 'multiple model' syllogism (Fig. 2). While the relationship between the components of the second syllogism differs across 2a and 2b, the original conclusion continues to hold in both cases.

Mental model theory predicts that resolving syllogism B will be more difficult as a result of the multiple ways in which the components of the syllogism can be represented. This increase in difficulty is measured by the percentage of participants who correctly classify the syllogism as valid or

not and by an increase in the amount of time required to come to such a decision (Byrne & Johnson-Laird 1989). Indeed, only 55% of reasoners given this multiple model argument correctly classify it as valid. So what makes this syllogism, and multiple model syllogisms in general, so much harder to evaluate than single model syllogisms?

Past research provides two possible answers to the above question. The first answer relates to the number of spatial representations that the reasoner must analyze. Specifically, mental model theory postulates that multiple model problems are more difficult to resolve because the premises allow for the creation of alternative models that must be mentally constructed and considered before a conclusion regarding validity can be reached (Johnson-Laird 1995). This process takes time and loads on cognitive capacity, which leads to mistakes and inefficiencies in reasoning. Importantly, on this account, both single and multiple model problems are resolved using similar cognitive processes defined over visuospatial representations (Johnson-Laird 1995).

A second possible explanation is that different types of cognitive processes are engaged when resolving multiple and single model arguments. It has been hypothesized that linguistic, in addition to visuospatial processes, are activated when participants resolve more difficult (i.e., multiple model) reasoning items. Mani & Johnson-Laird (1982) consider this possibility based on data indicating that while subjects remember the gist of single model descriptions better, they have better memory for the verbatim details of multiple model descriptions. They explain these results by postulating the existence of two different types of encoding (i.e., representations): Propositional (i.e., linguistic) and analogical (i.e., visuospatial). According to this account, reasoning is a multi-step process. The first step consists of forming a loose and superficial linguistic/propositional representation of each sentence. This surface representation is sufficient for the encoding of verbatim information. The second step involves the construction of a visuospatial mental model that is consistent with the perceptual layout of the linguistic/propositional representation. Although linguistic/propositional representations are necessary for the formation of mental models, in all likelihood they are discarded after mental models are formed. The inference itself is defined strictly over the visuospatial model.

An alternative view for accommodating this data is to give the linguistic representations a more central role in the reasoning process, resulting in a "mixed model" account that involves a combination of linguistic and visuospatial processes (Van der Henst & Schaeken in press). When resolving single model problems, the reasoner may construct a spatial mental model of the relationship between the premises and conclusion. However, when the reasoning items become more complex (i.e., multiple model problems), the reasoner augments the mental construction and evaluation of

spatial models with linguistic representations and inferential processes. According to this mixed visuospatial/linguistic approach, effects linked to a linguistic/propositional representation should occur more frequently with multiple model problems than with single model problems (Van der Henst & Schaeken in press).

Behavioural data has not been able to distinguish between these two explanations. An examination of brain activation while subjects solve single and multiple model problems provides another source of data to address the issue. If it is indeed the case that reasoning involves only the visuospatial system, then one would expect the involvement of the right hemisphere (Johnson-Laird 1995), or more accurately, occipital and parietal systems. Furthermore, it follows that as problems become more difficult (i.e., require the evaluation of multiple mental models), increasing task demands will result in greater activation in visuospatial systems.

However, if it is the case that linguistic representations play a significant role in the reasoning process, particularly in the case of multiple model problems, then one would expect greater involvement of left hemisphere frontal temporal systems in such trials. This prediction is consistent with much of the work in neuropsychology and cognitive psychology that stresses the importance and necessity of the left hemisphere for higher cognition including reasoning and problem solving. For example, tests of intelligence and general cognitive ability, such as the Scholastic Aptitude Test (SAT), Raven Matrices, and various vocabulary and reading comprehension tests, are highly correlated with logical reasoning (Stanovich & West 2000, Stanovich et al. 2004). These general cognitive tests are associated with activation in the left hemisphere, and specifically left lateral and dorso-lateral prefrontal cortex (Smith & Jonides 1997). Furthermore, it is reported that even after commissurotomy, where the two hemispheres are separated from each other, the left hemisphere continues to function at or close to preoperative levels (Gazzaniga 1970, 1989, 1995). The right hemisphere, on the other hand, is seriously impaired on cognitive tasks, especially in its ability to reason and solve problems (Gazzaniga 1970, 1989, 1995).

A series of neuroimaging and patient studies of human reasoning by various groups (Goel et al. 1998, Knauff et al. 2003, Langdon & Warrington 2000, Wharton et al. 2000) have consistently reported left hemisphere dominance for logical reasoning, while a series of studies by Goel and colleagues suggest that multiple neural pathways underlie human reasoning (Goel et al. 2000, Goel & Dolan 2001, 2003, 2004). According to Goel and colleagues, and consistent with a visuospatial account, a bilateral parietal (left > right) system is activated when processing unfamiliar, nonconceptual or incoherent material (e.g., All P are B; All C are P; $\therefore$ All C are B), while (consistent with a linguistic account) a left frontal-temporal lin-

guistic system is activated when processing familiar, conceptually coherent material (e.g., All dogs are pets; All poodles are dogs; ∴ All poodles are pets).

However, the neural basis underlying the resolution of single versus multiple model syllogisms has yet to be examined. To test competing hypotheses regarding the relative role of visuospatial and linguistic system involvement in reasoning, particularly in response to increasing number of mental models, we reorganized published archival data into single and multiple model problems, and reanalyzed it along this dimension.

2. Method

We conducted a reanalysis of data that was collected for an earlier reasoning study (Goel & Dolan 2003). The methods described here are those utilized in that study.

2.1. Subjects

We scanned 14 right-handed normal subjects using event-related fMRI, which indexes task-related activity, while the subjects engaged in deductive reasoning. Seven right-handed males and seven right-handed females with a mean age of 30.8 years (SD = 4.3) and a mean education level of 16.8 years (SD = 2.0) volunteered to participate in the study. All subjects gave informed consent and the study was approved by the Joint National Hospital for Neurology and Neurosurgery/Institute of Neurology Ethics Committee (UCL London).

2.2. Stimuli

We reorganized the stimuli in the original study (Goel & Dolan 2003) to look at performance on valid single model (n=21) and multiple model (n=19) trials and 20 relevant baseline trials. The non-reasoning or baseline condition trials were generated by randomly taking approximately half of both the single and multiple syllogisms and switching around the third sentence such that the three sentences did not constitute arguments. All sentences used in the study were grammatical, meaningful, and matched for length across conditions. As such, we ended up with a 2 x 2 study design with difficulty (single versus multiple) and task (reasoning versus baseline) as the two variables of interest (see Fig. 3).

Stimuli from all conditions were presented randomly in an event-related design (Fig. 4). A "*" indicated the start of a trial at 0 s. The sentences

		Difficulty	
		Single Model	**Multiple Model**
Task	**Reasoning**	All California snails are amphibians. No amphibians can sing. No California snails can sing. (21)	No Cambodian lizards are make-belief. Some Cambodian lizards are dragons. Some dragons are not make-belief. (19)
	Baseline	All California snails are amphibians. No amphibians can sing. All marathon runners are healthy. (12)	No Cambodian lizards are make-belief. Some Cambodian lizards are dragons. Some parents are not respected. (8)

Total N (syllogisms) = 60

Numbers in brackets refer to the number of syllogisms in each cell

Fig. 3. Overall design of study with sample stimuli

appeared on a screen one at a time with the first sentence appearing at 500 ms, the second at 3500 ms, and the last sentence at 6500 ms. All sentences remained on the screen until the end of the trial. The length of trials varied from 10.25 to 14.35 s, leaving subjects 3.75 - 7.85 s. to respond. The task in all conditions was the same. Subjects were required to determine whether the conclusion followed logically from the premises (i.e., whether the argument was valid). Participants responded by pressing a button on a keypad after the appearance of the last sentence.

In reasoning trials where the three sentences constituted an argument, participants had to determine the validity of the argument. In baseline trials, where the third sentence was unrelated to the first two, participants would begin to construct a representation of the problem, but could disengage and respond "no" with the appearance of the third unrelated sentence. Participants were instructed to respond as quickly as possible and move to the next trial if the stimuli advanced before they could respond. Participants reviewed example stimuli from each condition prior to being scanned to ensure that they understood the task. Participants were not given feedback about their performance during the experiment.

2.3. FMRI SCANNING TECHNIQUE

A 2T Siemens VISION system (Siemens, Erlangen, Germany) was used to acquire T1 anatomical volume images (1x1x1.5 mm voxels) and 48 T2*-weighted echoplanar images (64x64 3x3 mm pixels, TE = 40 ms) sensitive to blood oxygenation level dependent (BOLD) contrast. Echoplanar images (1.8 mm thick) were acquired axially every 3-mm, positioned to cover the

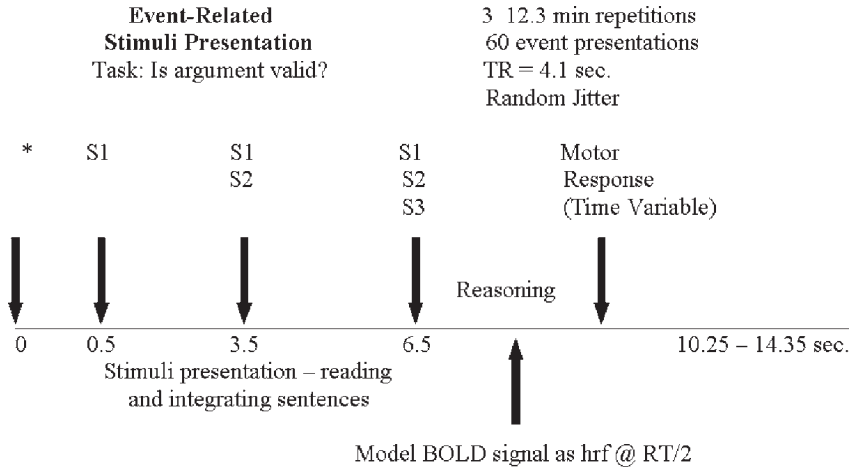


Fig. 4. Stimuli presentation

whole brain. Data were recorded during a single acquisition period. A total of 558 volume images were acquired over three sessions (186 volumes per session) with a repetition time (TR) of 4.1 s/volume. The first six volumes in each session were discarded (leaving 180 volumes per session) to allow for T1 equilibration effects.

Trials from all conditions were randomly presented in a single-event design. The mean trial time was 12300 ± 2050 ms (TR) with a random jitter. Trials thus varied from 10.25 to 14.35 s. There were 60 event presentations during a session for a total of 180 over three sessions. Each session lasted 12.3 min. The scanner was synchronized with the presentation of all trials in each session.

2.4. Data analysis

Imaging data were analyzed using Statistical Parametric Mapping (SPM 2) (Friston et al. 1995). All volumes were spatially realigned to the first volume (head movement was <2 mm in all cases) and temporally realigned to the AC-PC slice, to account for different sampling times of different slices. A mean image created from the realigned volumes was co-registered with the structural T1 volume and the structural volumes spatially normalized to the Montreal Neurological Institute brain template (Evans et al. 1993) using non-linear basis functions (Ashburner & Friston 1999). The derived spatial transformation was then applied to the realigned T2* volumes, which were finally spatially smoothed with a 12 mm FWHM isotropic Gaussian kernel (in order to make comparisons across subjects and to permit application of

random field theory for corrected statistical inference (Worsley & Friston 1995). The resulting time series across each voxel were high-pass filtered with a cut-off of 128 s, using cosine functions to remove section-specific low frequency drifts in the BOLD signal. Global means were normalized by proportional scaling to a Grand Mean of 100, and the time series temporally smoothed with a canonical hemodynamic response function to swamp small temporal autocorrelations with a known filter.

Condition effects at each voxel were estimated according to the general linear model and regionally specific effects compared using linear contrasts. Each contrast produced a statistical parametric map of the t-statistic for each voxel, which was subsequently transformed to a unit normal Z-distribution. The BOLD signal was modeled as a HRF at the midway point between the presentation of the third sentence and the motor response on a trial-by-trial basis. The presentations of all three sentences as well as the motor response were modeled out in the analysis. All results presented survived a significance level of p=.005 uncorrected.

3. Results

Overall behavioural results were analyzed in SPSS using repeated-measures ANOVA (single vs. multiple vs. baseline). This analysis revealed a significant difference in accuracy between the conditions, $F(2,12)$=6.59, p=.01. Further Bonferonni-corrected paired-samples t-test post-hoc analyses revealed a significant difference in accuracy between the single and baseline syllogisms, $t(1,13)$=3.65, p<.01 as well as the multiple and baseline syllogisms, $t(1,13)$=3.42, p<.01. There was no significant difference in accuracy or reaction time between the single and multiple model syllogisms.

Only those syllogisms that were answered correctly by participants were included in the imaging analysis. This step was taken to reduce variability in the imaging results as accurate responses indicate that participants were actually engaged in the reasoning task. The main effect of reasoning was determined by comparing all reasoning trials to all baseline trials [(single and multiple models) - baseline trials]. This analysis revealed activation in a largely left hemisphere system involving left lateral and dorso-lateral prefrontal cortex, left superior parietal lobule, left middle temporal lobe, primary visual cortex, precuneus, medial dorsal prefrontal cortex, and right lateral prefrontal cortex (Fig. 5).

To isolate brain regions associated with reasoning about multiple model syllogisms (but not single model syllogisms) and single model syllogisms (but not multiple model syllogisms), we directly compared the two conditions. A comparison of single model syllogisms versus multiple model

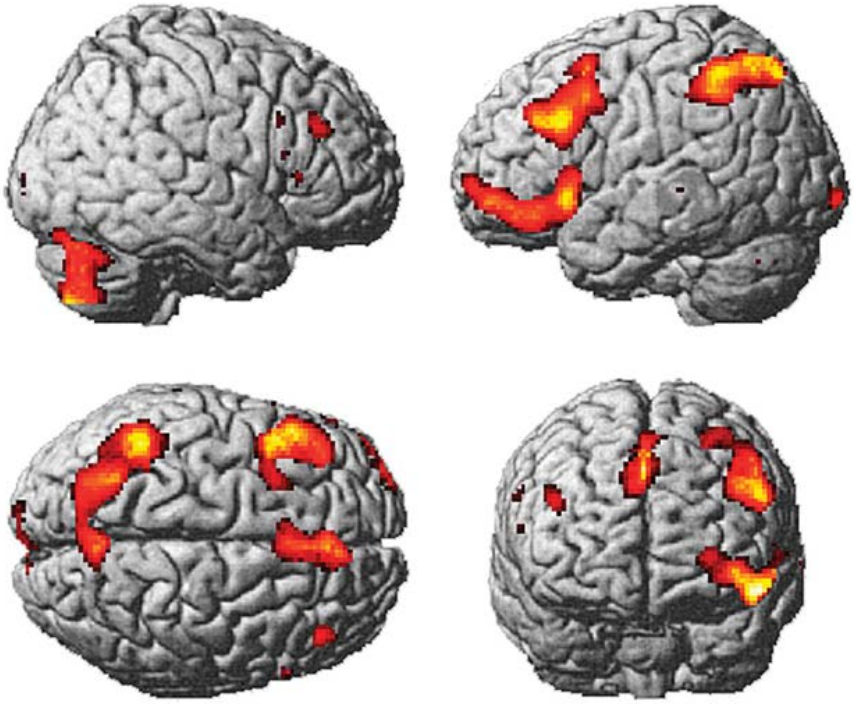


Fig. 5. Areas of activation for all reasoning items—baseline

syllogisms (masked inclusively by the main effect of reasoning) revealed activation in a largely left hemisphere network consisting of medial dorsal prefrontal cortex (BA 8) (8, 38, 44; $z = 3.28$) and left superior parietal lobule (BA 7) (- 24, - 56, 40; $z = 3.59$) (Fig. 6).

The reverse comparison of multiple model syllogisms versus single model syllogisms (masked inclusively by the main effect of reasoning), revealed activation in an exclusively left hemisphere network consisting of precuneus (BA 7) (-8, -76, 54; $z = 3.27$), inferior parietal lobule (BA 40) (- 38, - 70, 52; $z = 2.71$), superior temporal lobe (BA 21/22) (-66, -34, 4; $z = 3.45$), and lateral PFC (BA 47) (-50, 40, -14: $z = 2.99$) (Fig. 7).

4. Discussion

The behavioral results of the main effect of reasoning in the present study indicated that participants were engaged in the reasoning task and further analyses were warranted. The imaging results of the main effect of reasoning replicated previous studies in which reasoning about items with

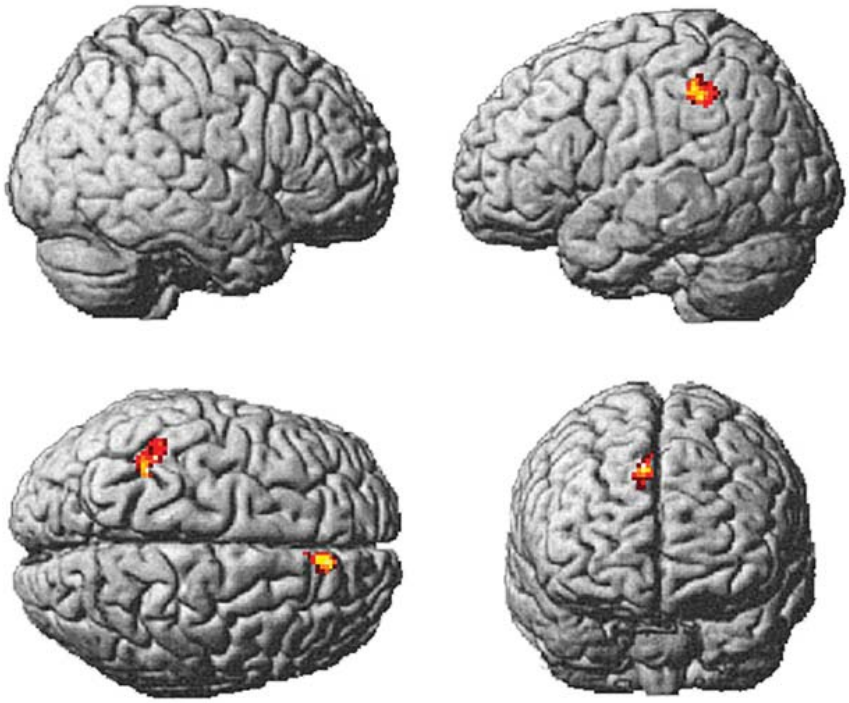


Fig. 6. Areas of activation for single model—multiple model

familiar, conceptually coherent material activated a left frontal-temporal language-based "heuristic" system (Goel 2003, Goel et al. 2000, Goel & Dolan 2001, 2003, 2004, Knauff et al. 2003).

The single model versus multiple model comparison revealed activation in dorsal medial PFC as well as left superior parietal cortex. We did not observe significant right hemisphere activation for single model problems, but we did observe activation in visuospatial areas in the left hemisphere. These results are consistent with greater involvement of the visuospatial system in the resolution of single model syllogisms.

In contrast, the multiple model versus single model comparison revealed activation in both left superior parietal lobule and left frontal and temporal language areas. This activation of areas implicated in both linguistic and visuospatial processing is of particular interest. It supports the position that the difference between resolving multiple and single model problems is not one of just greater visuospatial and working memory resources but rather increased involvement of the language system. On the surface, this result could be consistent with either the Mani & Johnson-Laird (1982) (i.e. superficial linguistic encoding which is not part of the inference pro-

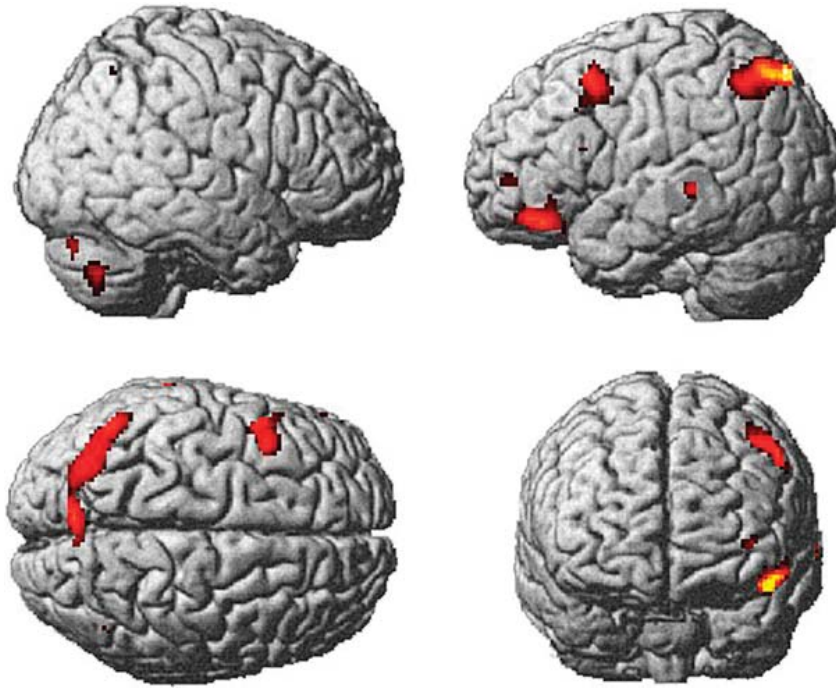


Fig. 7. Areas of activation for multiple model—single model

cess) or the Van der Henst & Schaeken (in press) (linguistic encodings play a critical role during inference) positions discussed above. However, the manner in which we have conducted our analysis supports the latter rather than the former account. Specifically, the BOLD signal in our study was modeled as a hemodynamic response function at the midway point between the presentation of the third premise and participant's motor response. The encoding of all three sentences and the participant's motor response were modeled as events of no interest. As such, if language-related areas are only involved in the first step of a multi-step model-building process, and are 'discarded' prior to the inference step (Mani & Johnson-Laird 1982), no language-related areas of activation should have been observed in our results. By contrast, if language-related areas are involved in the actual inference process (at least in reasoning about multiple model problems), as suggested by a mixed model approach (Van der Henst & Schaeken in press), language-related areas should be activated in our results. This is exactly what we found.

In summary, there is considerable evidence from both the lesion and neuroimaging literature that both visuospatial and linguistic processes play an important role in logical reasoning (Goel et al. 1998, Langdon & Warring-

ton 2000, Wharton et al. 2000). What the current study adds to these data is that as reasoning problems become more difficult (i.e. move from single to multiple models) there is increased activation in left hemisphere linguistic systems, suggesting that reasoners are augmenting any visuospatial model that they may have constructed with linguistic representations, and that these representations play an important role in the inference process.

References

Ashburner, J. & Friston, K. J. (1999), 'Nonlinear spatial normalization using basis functions', *Human Brain Mapping* **7**(4), 254–266.

Byrne, R. M. J. & Johnson-Laird, P. M. (1989), 'Spatial reasoning', *Journal of Memory and Language* **28**, 564–575.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L. & Peters, T. M. (1993), 3D statistical neuroanatomical models from 305 MRI volumes, *in* 'Proceedings of the IEEE-Nuclear Science Symposium and Medical Imaging Conference', IEEE Press, Piscataway, NY, pp. 1813–1817.

Friston, K., Holmes, A., Worsley, K., Poline, J. B., Frith, C. & Frackowiak, R. (1995), 'Statistical parametric maps in functional imaging: A general approach', *Human Brain Mapping* **2**, 189–210.

Gazzaniga, M. S. (1970), *The bisected brain*, Appleton-Century-Crofts, New York, NY.

Gazzaniga, M. S. (1989), 'Organization of the human brain', *Science* **245**, 947–952.

Gazzaniga, M. S. (1995), 'Principles of human brain organization derived from split-brain studies', *Neuron* **14**, 217–228.

Goel, V. (2003), 'Evidence for dual neural pathways for syllogistic reasoning', *Psychologica* **32**, 301–309.

Goel, V., Buchel, C., Frith, C. & Dolan, R. J. (2000), 'Dissociation of mechanisms underlying syllogistic reasoning', *Neuroimage* **12**(5), 504–514.

Goel, V. & Dolan, R. J. (2001), 'Functional neuroanatomy of three-term relational reasoning', *Neuropsychologia* **39**, 901–909.

Goel, V. & Dolan, R. J. (2003), 'Explaining modulation of reasoning by belief', *Cognition* **87**, B11–B22.

Goel, V. & Dolan, R. J. (2004), 'Differential involvement of the left prefrontal cortex in inductive and deductive reasoning', *Cognition* **93**, B109–B121.

Goel, V., Gold, B., Kapur, S. & Houle, S. (1998), 'Neuroanatomical correlates of human reasoning', *Journal of Cognitive Neuroscience* **10**(3), 293–302.

Johnson-Laird, P. N. (1995), Mental models, deductive reasoning, and the brain, *in* M. S. Gazzaniga, ed., 'The Cognitive Neurosciences', MIT Press, Cambridge, MA, pp. 999–1008.

Knauff, M., Fangmeier, T., Ruff, C. C. & Johnson-Laird, P. N. (2003), 'Reasoning, models, and images: Behavioral measures and cortical activity', *Journal of Cognitive Neuroscience* **15**, 559–573.

Langdon, D. & Warrington, E. K. (2000), 'The role of the left hemisphere in verbal and spatial reasoning tasks', *Cortex* **36**(5), 691–702.

Mani, K. & Johnson-Laird, P. N. (1982), 'The mental representation of spatial descriptions', *Memory & Cognition* **10**(2), 181–187.

Smith, E. E. & Jonides, J. (1997), 'Working memory: A view from neuroimaging', *Cognitive Psychology* **33**(1), 5–42.

Stanovich, K. E., Sa, W. & West, R. F. (2004), Individual differences in thinking, reasoning, and decision-making, *in* J. P. Leighton & R. J. Sternberg, eds, 'The Nature of Reasoning', Cambridge University Press, Cambridge.

Stanovich, K. E. & West, R. F. (2000), 'Individual differences in reasoning: Implications for the rationality debate?', *Behavioural and Brain Sciences* **23**, 645–726.

Van der Henst, J. B. & Schaeken, W. (in press), 'The wording of conclusions in relational reasoning', *Cognition* (in press).

Wharton, C. M., Grafman, J., Flitman, S. S., Hansen, E. K., Brauner, J., Marks, A. & Honda, M. (2000), 'Toward neuroanatomical models of analogy: A positron emission tomography study of analogical mapping', *Cognitive Psychology* **40**(3), 173–197.

Worsley, K. J. & Friston, K. J. (1995), 'Analysis of fMRI time-series revisited — again', *NeuroImage* **2**, 173–181.

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

A Neuro-Cognitive Theory of Relational Reasoning with Mental Models and Visual Images

Markus Knauff[1]

Max-Planck-Institute for Biological Cybernetics, Tübingen and
Center for Cognitive Science, Universität Freiburg[2]

Abstract
Recent brain imaging studies have provided evidence that the parietal cortex plays a key role in reasoning based on mental models, which are supposed to be of abstract spatial nature. However, these studies have also shown concurrent activation of visual association cortices which have often been interpreted as evidence for the role of visual mental imagery in reasoning. The present chapter resolves these inconsistencies. I argue that visual brain areas are only involved if the problem information is easy to visualize and when this information must be processed and maintained in visual working memory. A regular reasoning process, however, does not involve visual images but more abstract spatial representations—spatial mental models—held in parietal cortices. Only these spatial representations are crucial for the genuine reasoning processes.

[1] The author of this paper is supported by a Heisenberg Award from the Deutsche Forschungsgemeinschaft (DFG) and by grants within the Transregional Collaborative Research Center on Spatial Cognition (SFB/TR 8). I thank Kristen Drake, Vinod Goel, Kai Vogeley, Gottfried Vosgerau, André Vandierendonck, and Lara Webber for many helpful comments on an earlier version of this paper.
[2] E-mail: markus.knauff@tuebingen.mpg.de

The old lady with the yellow hat stands to the right of the man with big ears.
The man with big ears stands to the right of the constable.
Does the old lady with the yellow hat stand to the left or to the right of the constable?

Individuals often say that they reason about such problems by forming a mental picture in their "mind's eye" and then look at this picture to find new information. Yet, is the experienced immediacy of visual imagery related to the underlying "reality" of mental representations and processes? And, why does reasoning seem inextricably linked with seeing in the "mind's eye"? Not only non-psychologists, but also many cognitive psychologists have claimed that reasoning is strongly linked to imagination and thus tried to explicate how mental imagery and reasoning are interconnected (e.g. De Soto et al. 1965, Kosslyn 1994). There are, however, also reasons to be skeptical concerning the role of visual mental images in reasoning. For instance, if reasoning relies on visual imagination then problems that are easy to visualize should be easier to solve than non-visual problems. The problem above, for instance, should be easier than the formally equivalent problem:

A is smarter than B.
B is smarter than C.
Is A smarter than C?

In both problems, new information can be inferred from what is already given. Several researchers varied the imageability of such reasoning problems but did not find any differences between problems that are easy or difficult to visualize (e.g. Johnson-Laird et al. 1989). Neuroimaging studies sometimes find neural activity in vision-related brain areas during reasoning with such problems, and sometimes no such activity is found. Moreover, computational systems of human reasoning show that human reasoning performance can be properly reconstructed without visual images (Ragni et al. 2005, Schlieder 1999, Schlieder & Berendt 1998). So what really happens in our brains if we subjectively experience visual mental images during reasoning? In this chapter, I argue that the same sort of *spatially organized mental models* underlie reasoning and that these models are not to be identified with visual images. We might "see" a visual image for the first problem but not for the second. However, what matters is not our subjective experience, but rather what is processed by our cognitive system. The chapter starts with a brief overview of previous findings on reasoning and mental imagery. Then it reports a number of neuroimaging studies (partly coming from our own lab) that explored the involvement of visual brain areas in reasoning. Then I report a recent event-related fMRI study that for the first time disentangles the neuro-cognitive subprocesses underlying different stages in the reasoning process, and at the same time overcomes the

potential visual confound in the previous studies on the neuronal basis of human reasoning. Based on these findings and several behavioral results I propose a neuro-cognitive *three-stage-theory* of reasoning with mental models and visual images. While many studies have implied that visual images play a key role in the reasoning process, in this account visual brain areas are only involved if the problem information is easy to visualize and when this information must be processed and maintained in visual working memory. A regular reasoning process, however, does not involve visual images but more abstract spatial representations—spatial mental models—held in parietal cortices. Only these spatial representations are crucial for the genuine reasoning processes. If, however, the spatial information must be retrieved from a visual image in order to construct the appropriate spatial mental model (as in the problem with the drolly looking folks) additional processes come into play and can even impede the process of reasoning.

1. Historical remarks and behavioral findings

During the early decades of the last century, a fierce academic debate about the role of images in human cognition took place in German psychology. Although the functions of the sensory systems were still of great interest to psychologists, another particular area of attention was now the role of visual imagery in thinking, reasoning, and problem solving. On the one hand, in 1910, Cheves Perky discovered that mental imagery supports visual perception and that people often merge mental images and what is actually seen. In other words, visual imaginations can be so similar to real perceptions that they can be mistaken for the latter (Perky 1910). On the other hand, in particular the "Würzburger Schule" promoted the assumption that thinking is possible without imagination. The claim was supported in an experiment by Karl Bühler, who asked participants, for instance, "Does a man have the right to marry the sister of his widow?" and afterwards asked them what had happened in their mind. Not one of the participants reported experiencing visual images. From his findings, Bühler concluded that thinking is possible without seeing in the mind's eye (Bühler 1909). However, other authors criticized the idiosyncrasy of Bühler's problems and for a long period of time, for most researchers it was a matter of fact that thinking calls for "imagination" in the literal sense—that is, the activity of envisaging objects and scenes in their absence (e.g. Titchener 1909).

Later, in Anglo-American psychology, publications on mental imagery engendered much controversy. Cognitive psychologists avoided the concept of imagery, given the harsh criticism it had received from behaviorists (Wat-

son 1913). In contemporary psychology, however, a wide range of evidence is compatible with the assumption that imagery is a vital part of human cognition, including the well-known studies of mental rotation, the mental scanning of images (cf. Kosslyn 1980, Shepard & Cooper 1982), and studies on the relationship between imagery and creative problem-solving, suggesting that visualization facilitates innovative solutions (Suler & Riziello 1987, Antonietti 1991, recent results in: Denis et al. 2001). Moreover, subsequent to the well-known imagery debate in the 1980s (overview in: Block 1981, Tye 1991), the majority of cognitive researchers agree on the assumption that cognitive processes can rely on a number of different representational formats.

Starting in the 1960s, cognitive psychologists also began to explore the role of visual images in *relational* reasoning. The two problems above are examples of such inferences. In the psychology of reasoning, such problems are called *transitive inferences*, *linear syllogisms*, or *three-term-series problems* (Johnson-Laird 1972, Sternberg 1980). The problem information is given by the two statements which are called *premises*, and the task is to find a *conclusion* that necessarily (logically) follows from these premises. Adding further premises, changing the order of premises and terms, etc., can result in more complex problems (overviews can be found in Evans et al. 1993, Manktelow 1999).

A pioneering reasoning study was carried out by De Soto et al. (1965), who argued that reasoners represent the entities of a relational reasoning problem as a mental image and then "read off" the conclusion by inspecting the image. Huttenlocher (1968) also argued that reasoners imagine an analogous physical arrangement of objects in order to cope with reasoning problems. Moreover, other authors report that reasoning is easier with problems that are easy to envisage than with problems that are hard to envisage (e.g. Shaver et al. 1975, Clement & Falmagne 1986). However, several studies have failed to detect any effect of imageability on reasoning. Johnson-Laird et al. (1989), for instance, examined reasoning with relations that differed in imageability—equal in height, in the same place as, and related to (in the sense of kinship)—and did not find any effect on reasoning accuracy. Newstead et al. (1986) reported a similar result, and Sternberg (1980) did not find any reliable correlation between scores on the imageability items of IQ-tests and reasoning ability. Overall, for a long time the results from many behavioral studies have been inconclusive and have left many questions unresolved.

2. Results from neuroimaging

With the development of new brain imaging methods the debate shifted from the behavioral findings towards the question of how reasoning and mental imagery is biologically realized in the human brain. Broadly speaking, the occipital lobe processes visual information. However, it is not only responsible for visual perception, but also contains association areas and appears to help in the visual recognition of objects and shapes. The occipital cortex can be divided into the primary visual cortex, also referred to as striate cortex or, functionally as V1, and to the visual association areas, also called the extrastriate cortex, or V2, V3, V4. The primary visual cortex receives visual input from the retina and is topographically organized, meaning that neighboring neurons have receptive fields in neighboring parts of the visual field. According to the cytoarchitectonic map of Brodmann (1909), this region is called Brodmann's area (BA) 17. The visual cortices have been frequently related to visual mental imagery. For instance, patients who are blind in one side of the visual field are also unaware of objects on that side when imagining a visual scene. If the patient turns the mental image around so that they had to "look" at the image from the opposite direction, they reported objects on the other side and ignored those which they had previously reported "seeing" (Mellet et al. 1998).

The strictest form of imagery theories has been elaborated on in the influential book by Kosslyn (1994). In this book, Kosslyn claims that during mental imagery the geometrical information of remembered objects and scenes are processed in the primary visual cortex. Consequently, one of the central research issues on imagery is whether the primary visual cortex and nearby cortical areas are activated by visual mental imagery. Indeed, this assumption is supported by a series of studies by Kosslyn and his colleagues, who found increased blood flow in BA 17 during mental imagery of letters (Kosslyn et al. 1993) and objects in different sizes (Kosslyn et al. 1997). Moreover, if participants imagined a letter, the larger letters activated a larger region of V1 while the smaller letters activated a smaller region (Kosslyn et al. 1993). Additional support for the strong imagery theory comes from studies by Kosslyn et al. (1999), Sabbah et al. (1995), and Chen et al. (1998).

More moderate approaches to visual mental imagery are related to the complete ventral pathway. Beyond the striate cortex, the ventral pathway, or "what" system, comprises parts of the temporal lobes (Ungerleider & Mishkin 1982). The most important areas are the inferior temporal (IT) cortex that typically responds to properties of objects, such as shape, texture and color. The anterior parts of the system processes information in a visual code and cannot be assessed by other modalities—hence, the system

is modality-specific. The main function of the system is to identify objects, i.e., compare stored objects with the object that is viewed. However, this pathway can also run in the opposite direction so that visual images can be generated top-down from memories.

Outside the occipital areas, the dorsal pathway, or "where" system, comprises parts of the two parietal lobes. They contain the primary sensory cortex which controls sensation and large association areas. The posterior parietal cortex (PPC) and the precuneus are considered as areas that combine information from different sensory modalities to form a cognitive representation of space. Although these areas have diverse functions and use a variety of sensory modalities, they are all responsible for processing information about spatial relationships (Andersen 1997).

The frontal cortex is involved in planning, problem solving, selective attention, and many other higher cognitive functions (including social cognition and emotion). The anterior (front) portion of the frontal lobe is called the prefrontal cortex. It is involved in executive processes in working memory and typically implicated when several pieces of information in working memory need to be monitored and manipulated. A related function is that the region underlies the integration of multiple relations. Waltz et al. (1999), for instance, showed that patients with damage to the prefrontal cortex were strongly impaired in any sort of reasoning calling for the integration of relations, whereas they performed normally in episodic and semantic memory tasks.

Early brain imaging studies on reasoning found little evidence that visual brain areas (in occipital cortex) are involved in reasoning (Goel et al. 1997, 1998). Then, however, an increasing number of studies reported activity in primary and secondary visual areas when participants were engaged in reasoning problems. This, for instance, was the case in a study by Goel et al. (2000) in which the volunteers had to solve different kinds of relational inferences. Moreover, Knauff et al. (2000) studied relational and conditional inferences that were presented acoustically via headphones to the participants (to avoid a confounding of mental imagery and visual perception). In this study, both types of reasoning problems resulted in activity in a bilateral occipitoparietal-frontal network distributed over parts of the prefrontal cortex and the cingulate gyrus, the inferior and superior parietal cortex, the precuneus, and the visual association cortex. Similar results have been reported in Ruff et al. (2003). Here, we scanned the brain activity of our participants and also measured their visuo-spatial ability with a well-known subset of tasks from an intelligence inventory. Interestingly, the brain activation was significantly modulated by the participants' visuo-spatial skill. The higher the participants' visuo-spatial skill, the better their reasoning performance, and the less activation was present in visual association areas during reasoning. This pattern conforms with recent findings on the

effects of skill level on neuronal activity. Accordingly, the reasoning problems seemed to have placed less demand on the visuo-spatial processing resources of participants with high skill levels, so that less activity in the relevant cortical regions was required.

3. Disentangling visual and spatial processing in reasoning

Studies from the literature and our earlier findings provide informal evidence that reasoning is occasionally accompanied by visual mental imagery. Alas, these studies were not designed to determine the exact role of visual images in reasoning and thus examined the brain activation during the whole reasoning process in a blocked fashion (e.g. Knauff et al. 2002) or just compared the neuronal processes during the conclusion of the reasoning problem with the presentation of irrelevant control sentences (e.g. Goel et al. 2000). In both paradigms it is impossible to determine whether the activity in occipital brain areas pointing to the employment of visual mental imagery is associated with the processing of premises, their maintenance in working memory, or with the actual reasoning process. Reasoning-related processes during different stages of problem processing and other cognitive processes are inseparably mixed. To overcome these disadvantages, our group recently conducted an fMRI study to disentangle the neuro-cognitive subprocesses underlying the different stages in the reasoning process and at the same time to avoid potential confounds in the previous studies on the neuronal basis of imagery and reasoning. In this study, we scanned the brains of our participants while they solved relational reasoning problems (Fangmeier et al. in press, Knauff, Fangmeier, Ruff & Sloutsky 2005). Since we aimed at keeping apart the pure reasoning process from the maintenance of information in working memory, in a second group of tasks participants had to simply keep the premises of the identical problems in working memory without making inferences. To avoid the need to read the premises and conclusions we replaced the sentences with graphical arrangements describing the spatial relations between three objects. The reasoning problems contained two premises and a conclusion and the participants had to decide whether the conclusion logically (necessarily) followed from the premises. Here is a example of a reasoning task with a valid conclusion:

premise 1:	V	X
premise 2:	X	Z
conclusion:	V	Z

A sentential version of the given example would be: "V is to the left of X" (first premise) and "X is to the left of Z" (second premise). From these

premises it follows "V is to the left of Z" (conclusion). In the maintenance problems, the presentation of the two premises was the same as in the reasoning task, but the participants had to decide whether the term order of the third sentence was identical to one of the previous premises or not. Thus, no inference between the two premises had to be made. Moreover, the processing of the first premise, the second premise and the conclusion was time-locked to the presentation of the arrangements. Thus, we could examine the brain activity elicited by different stages of the reasoning process.

The results of this study are illustrated in Figure 1. The darker a region in the image is, the more cortical activity was measured. As can be seen from the foci of activation, we identified three distinct patterns of neuronal activation associated with three stages of the reasoning process. During the presentation of the first premise, reasoners had to process and maintain the spatial relation between the first two objects in working memory. During this stage we found two large bilateral clusters of activation in the vision-related occipito-temporal cortex (see Figure 1a). Then the participants needed to unify the second premise with the information from the first premise in order to construct an integrated representation of both premises. During this stage the two clusters in the occipito-temporal cortex and an additional cluster in the anterior prefrontal cortex (AFC) were activated. The latter cluster covered parts of the middle frontal (BA 10) and medial frontal gyrus (BA 32; see Fangmeier et al. 2005 for details). In the third stage participants had to inspect and manipulate this representation to draw a putative conclusion and to compare this conclusion with the displayed conclusion. They indicated by pressing a button whether the displayed conclusion is "True" or "False." Crucially, this stage activated clusters in the dorsolateral prefrontal cortex (DLPFC) and in the spatial areas of posterior parietal cortex, whereas vision-related activity in occipital cortex completely disappeared.

The contrasts between the reasoning and maintenance of premises were carried out to separate the pure reasoning process from the maintenance of information in working memory. It is critical to appreciate that the processing of the matched maintenance problems also proceeded in three stages, but that participants only had to remember the premises and match it with the presented third arrangement. They did not make any inferences. As also shown in Figure 1, the patterns of activity were similar only in the first stage but significantly differed from reasoning in the second and third stages. During the first stage of the maintenance problems we again found activity in the two large bilateral clusters in the vision-related occipito-temporal cortex that we also obtained during reasoning (compare Fig. 1a with 1d). In the second stage, which now required only premise maintenance but not integration, we again found similar activation in occipital areas, but

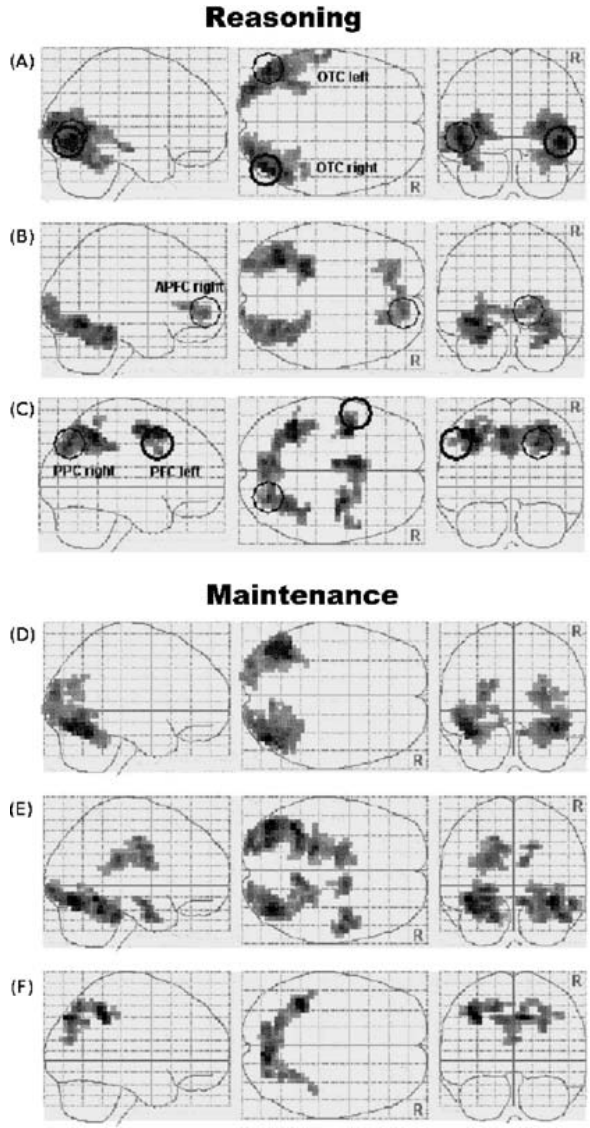


Fig. 1. Images representing differentially activated brain areas during the three stages of reasoning and maintenance. The brain is presented from three different perspectives. The clusters above indicate the activity for the reasoning tasks during (a) premise processing stage, (b) integration stage, (c) validation stage. The clusters below show the activity in the maintenance tasks during (d) premise processing stage, (e) premise maintenance stage, (f) validation stage. (from: Knauff, Fangmeier, Ruff & Sloutsky 2005; see text for details).

crucially no frontal activation (compare Figure 1b with 1e). Finally, during the third stage of the maintenance problems, there were significantly lower prefrontal activations and less extensive activation in space-related parietal areas than during the reasoning problems (compare Fig. 1c with 1f).

4. A neuro-cognitive three-stage theory of reasoning with mental models and visual images

As there is a many to many mapping between cortical regions and cognitive functions, neuropsychological data alone are too weak to formulate cognitive theories. However, if imaging data are consistent with behavioral findings this can provide strong support for a cognitive theory of human reasoning. The following sections are in the spirit of this connection between behavioral findings and neuropsychological results and thus employ both classes of experimental findings to introduce a neuro-cognitive theory of human (relational) reasoning that accounts for the different functions of visual *and* spatial representations in reasoning.

Take, for instance, the example at the beginning of this article. Reasoners might imagine three individuals—an old lady with a yellow hat, a man with the big ears, and a man in a constable's uniform—in a vivid visual image and think that they should use this image to find a relation not explicitly given in the premises. However, let us use a more neutral version of the problem to explain what could really happen during reasoning. Psychologists often use problems with tools, fruits, vegetables, etc., because they are easier for their participants to visualize and have less to do with their prior knowledge (Knauff, Jahn & Vosgerau 2005). So imagine for instance, the almost identical inference problem:

The hammer is to the right of the pliers.

The pliers are to the right of the screwdriver.

Does it follow that the hammer is to the right of the screwdriver?

The findings by Fangmeier et al. (in press) indicate that such inferences depend on three neuro-cognitive stages of thought. In the following, I refer to these stages as (1) *visual image construction*, (2) *image to model transformation*, and (3) *mental model processing*. I will show that this distinction is consistent with many behavioral findings.

Visual image construction. Our data show that this stage relies on neural processes in the occipito-temporal cortex that are known to be involved in visual mental imagery and visual working memory. The most reasonable account for this finding is that the processing of the first premise spontaneously elicits visual imagery. Reasoners seem to use their background knowledge to construct a *visual mental image* of the information from the

premise. They, for instance, imagine the tools lying on a table or on the floor of their garage. Two sorts of knowledge are needed for this visual image construction: knowledge about the visual features of the objects and knowledge referring to the meaning of the spatial expressions. The former is provided by the visual pathway that is known to run in two directions. Processing during perception begins with a retinotopic representation in the occipital cortex and progresses to memory representations of objects in areas of the temporal cortex. However, visual images also can be generated top-down from memories: Visual information stored in memory travels backwards from the temporal regions of the ventral pathway into the occipital cortex where it evokes a pattern of activity that is experienced as a mental image (Farah et al. 1988). For knowledge about spatial relations a similar mechanism exists. One of the best investigated areas of the brain is the posterior (back) part of the parietal cortex which receives projections from extrastriate visual areas and projects to areas associated with saccadic eye movements. In the present context, however, it is important that these areas of the dorsal pathway form a mental representation of space. During perception spatial relations are extracted from the retinotopic representations in the occipital cortex, and result in memory representations in the posterior parietal cortex (PPC). However, this spatial information can also be generated top-down from memory so that an object from the ventral pathway (e.g., the tools in the example) can be located in the visual image. The resulting visual image is structurally similar to a real visual perception and relies on similar brain functions. Like a visual percept, it might represent colors, shapes, and metrical distances. It probably can be rotated and scanned and it might have a limited resolution (cf. Kosslyn 1994, Johnson-Laird 1998). It is reasonable to assume that these representations of the premises are responsible for the experience of visual images during reasoning. Reasoners might be aware of the visual images, but they probably do not have conscious access to what is going on in the next steps of the inference.

Image to model transformation. The essential finding for this stage is the activity in the anterior prefrontal cortex (AFC). Neural computations in these areas seem to bridge the gap between the visual image of the premises and the third stage of reasoning, where vision-related activity in the occipital cortex completely disappears and is replaced by large activated clusters in spatial brain areas in the posterior parietal cortex (PPC). The most plausible explanation for this finding is that the actual reasoning is based on spatial representations and the visual images of the premises are not pertinent to the reasoning processes. Therefore, the spatial information must be *retrieved* from the visual image in order to construct the appropriate spatial mental model for making the inference. Thus, there must be a mechanism that transforms visual representations into spatial ones. The resulting spa-

tial representations might be, as many results suggest, *mental models* in the sense of Johnson-Laird (1983) and Johnson-Laird & Byrne (1991). Such models represent the information pertinent to reasoning by means of spatial relations. In inferential tasks, the resulting spatial representations are likely to exclude visual detail and to represent only the information relevant to the inference. They take the form of a representation that maintains the spatial relations between objects in a multi-dimensional array. According to model theory, such a spatial representation of the premises above could be the following:

screwdriver pliers hammer

There is substantial evidence to suggest that the anterior prefrontal cortex is involved in the processing of relations. Specifically, this area has been found to be involved in relational integration during reasoning or in considering multiple relations simultaneously (e.g. Waltz et al. 1999, Christoff et al. 2001). Relational integration appears to be a specific kind of mental computation that develops slowly in humans—as much as deductive reasoning ability does (cf. Evans et al. 1993). Moreover, the neural computation is strongly influenced by the number of relations that must be considered. Halford et al. (1998) distinguished three levels of complexity: in 0-relational problems, no relations need to be considered; in 1-relational problems, a single relation must be considered; and in 2-relational problems, two relations must be considered simultaneously and, thus, integrated. All problems from the fMRI studies reported here belong to the last group of problems because exactly two relations must be retrieved from the visual images. In the example above it is the relation between the hammer and the pliers and the relation between the screwdriver and the pliers. It is important to see that the third relation, namely that between the hammer and the screwdriver, does not need to be explicitly represented because it can be read off from the model. Moreover, it is essential to see that these processes are unlikely to be accessible to the conscious experience of the individual. The reasoner still just experiences the image of the premises.

Mental model processing. In the final stage, we found activations in the bilateral PPC and the dorsolateral prefrontal cortex (DLPFC). While the other two stages were basically concerned with the visual image and its transformation into a spatial model, this stage lies in the heart of reasoning. Now the spatial mental model must be processed by logical routines. The maintenance and handling of spatial representations is known to be managed by regions in the PPC. According to many studies, the PPC plays a crucial role in the processing of spatial information from different modalities (Burgess et al. 2001) and in the integration of sensory information from all senses into egocentric spatial representations (Andersen et al. 1997, Bushara et al. 1999, Colby & Duhamel 1996, Xing & Andersen 2000). Crucially, these areas are not exclusively dedicated to information coming

from visual perception. Several studies show that areas in the PPC bring spatial information from all perceptual systems into the same reference system. Another important finding is that of the laterality of the human PPC. Kosslyn et al. (1989) have shown that there are two different subsystems processing quantitative-metrical and qualitative-categorical spatial information (see also: Kosslyn et al. 1992). Metrical spatial information is that in which exact distances with respect to a continuous coordinate system are represented, and Kosslyn located this system in the right hemisphere. In contrast, categorical spatial information is that in which spatial relations between objects are represented qualitatively by discrete spatial concepts. Although these relations are presumably not represented in a language-based format, the concepts may correspond to verbal expressions such as left and right, above and below (Knauff 1999).

According to model theory, the spatial representation captures one situation that is possible, given that the premises are true (Johnson-Laird 1983, Johnson-Laird & Byrne 1991). Like a spatial diagram, the model's parts correspond to the parts of what it represents, and its structure corresponds to the structure of the reasoning problem (Johnson-Laird 2001). In other words, a mental model is a representation of objects and relations that constitutes a model (in the usual logical sense) of the premises given in the reasoning task. According to the model theory, reasoning with this model relies on processes that inspect and validate the model. The inspection yields new information that is not explicitly given in the premises and the validation checks whether a putative conclusion is actually true. As computational models suggest, the inspection process can be functionally described as a shift of a spatial focus that checks the cells of a spatial array and "knows" from the scan direction the relation between two objects in the array (Ragni et al. 2005, Schlieder & Berendt 1998).

In the present account, the model is represented in the neural tissue of the parietal cortex and the inspection and validation processes are controlled by computations in the PFC. It is very likely that reasoners are not aware of all of these processes, because deductive reasoning—like fundamental memory processes—has to be performed extremely fast and accurately, and must be sheltered from external disruptions. Nevertheless, the current account is suggested by many studies on cognitive control, characterizing sections of the PFC as typically involved when several pieces of information in working memory need to be monitored and manipulated (Petrides 2000). Moreover, patients with damage to the prefrontal cortex are strongly impaired on deductive (and inductive) reasoning tasks whenever these require the processing of relations (e.g. Waltz et al. 1999). Together with our present findings, this indicates that structures in the PFC and PPC strongly interact during reasoning. Parietal areas are concerned with the mental model itself and the PFC is responsible for controlling the inspection and manipulation of

this model. Normally, these processes work error-free and thus results in a valid conclusion, i.e., that, in the example above, the screwdriver is to the left of the hammer. Errors do occur, however, because reasoning performance is limited by the capacities of the systems, the misunderstanding of the premises, or the ambiguity of problems (Johnson-Laird & Byrne 1991, Evans et al. 1993, Manktelow 1999).

Although the account is not yet spelled out in all details, it resolves many inconsistencies in previous neuroimaging studies on reasoning. These studies have similarly implied that the parietal cortex may play a key role in reasoning based on mental models, which are supposed to be of abstract spatial nature. However, these studies have also shown concurrent activation of visual association cortices (Goel & Dolan 2001, Goel et al. 2000), which have often been interpreted as evidence for the role of visual mental imagery in reasoning (Ruff et al. 2003). The present account makes this role of images clearer. It shows, for the first time, that visual brain areas might be involved in premise processing and the construction of an initial visual image of the situation described in the premises. These processes, however, are not specific to reasoning, but primarily related to the comprehension of premises and their visual representation in working memory. The actual reasoning process then relies on more abstract spatial representations held in parietal cortices. Because initially a visual image had been constructed from the premises, the spatial information relevant for reasoning must be retrieved from this image in order to construct the appropriate spatial mental model for making the inference. The inspection and manipulation of these spatial mental models is crucial for subsequent processes and the supplementary activation in the DLPFC and AFC during reasoning indicates that further processes are exclusively devoted to the processing of relations and executive control processes. Individuals might be aware only of the visual images, but it is also possible that we do not have conscious access to the spatial representations and the processes that inspect and manipulate this representation, although they underlie our reasoning abilities.

5. Further evidence for the theory

The theory presented here relies on two major conjectures: Visual images are involved in the processing and maintenance of premises in working memory, but not in the actual reasoning process. And: The spatial relations from the premises must be integrated into one spatial mental representation—the mental model—in order to make the inference. This spatial model can then be further processed by logical routines that inspect and manipulate the model. Both assumptions are supported by further experimental

findings.

Conjecture 1: Visual images are involved in the processing and maintenance of premises in working memory. Support for this claim comes from two groups of studies. First, countless studies in the field of text comprehension have shown that visual representations are routinely and immediately activated during word and sentence comprehension. If individuals are asked to read texts but were given no instruction to form visual images they regularly experience visual images while reading (cf. Sadoski & Paivio 1994). Most of the explanation is more or less inspired by the well-known dual-coding theory in which cognition relies on two separate but interconnected systems: a verbal system for language and a nonverbal system that deals with visual images (Paivio 1971, 1986). Today, almost everybody in reading research has no doubt that mental imagery occurs as a spontaneous process in reading and that images have powerful effects on comprehension, recall, recognition, and the reception of the text (e.g. Glenberg 1997, Sadoski 1985, Sadoski & Paivio 1994, Stanfield & Zwaan 2001, Zwaan et al. 2002).

Evidence that visual images are primarily involved in the processing and maintenance of premises in working memory also comes from the comparison of reasoning and maintenance problems. An initial study has been conducted by Ruff et al. (2003) who examined the differences between both tasks in a blocked design. Interestingly, neuronal activations common to reasoning and maintenance were detected bilaterally in secondary visual cortices. This again indicates that the occipital activation patterns were not related to reasoning, but rather to the mere encoding and maintaining of premises in visual working memory. A second finding was that only reasoning led to more activation than maintenance bilaterally in the dorsolateral prefrontal cortex and in the anterior prefrontal cortex. As already mentioned, Waltz et al. (1999) showed that patients with damage to the prefrontal cortex were strongly impaired on deductive and inductive reasoning tasks whenever these required relational integration. Waltz et al. concluded that "postulating a neural system for integrating multiple relations provides an explanation of why a wide range of tasks, all of which depend on processing multiple relations simultaneously, are sensitive to prefrontal damage and activate DLPFC" (p.124). For the present account it is essential that relational integration is a vital part of reasoning with transitive inferences, while it is not required for solving the maintenance problems.

Conjecture 2: Premises during reasoning are integrated into one unified mental representation and this representation is inspected to find new information. This assumption is also supported by two groups of findings. The first is related to the work on relational integration and the connection between complexity and number of relations (Halford et al. 1998). Christoff

et al. (2001) tested the hypothesis that the process of relational integration is a component process of complex reasoning and that it recruits PFC. They examined brain activation during 0-relational, 1-relational, and 2-relational problem solving and found that PFC is more activated by 2- than by 1-relational problems and by 1-relational problems more than by 0-relational problems. This link between neural activity and the number of relations reflects that relations must be integrated into one unified representation and this is associated with processes of manipulating self-generated new information.

The second group of supporting studies is linked to mental models research. An important prediction of model theory is that the ease of reasoning is a function of the difficulty to integrate the information from the premises into a unified representation. Hence, Ehrlich & Johnson-Laird (1982) gave subjects the premises of a transitive inference in continuous (A r_1 B, B r_2 C, C r_3 D), semi-continuous (B r_2 C, C r_3 D, A r_1 B), and discontinuous (C r_3 D, A r_1 B, B r_2 C) premise orders (the letter r stands for a certain relation). Subjects had to infer the conclusion A r_4 D and the results showed that continuous order (37% error) is easier than discontinuous order (60% error), and there is no significant difference between continuous and semi-continuous (39% error) tasks. This finding is an effect of the difficulty of integrating the information from the premises into a unified representation because in the continuous and semi-continuous orders, it is possible to integrate the information of the first two premises into one representation—a mental model—at the outset, whereas when they are presented with the discontinuous order, subjects must wait for the third premise in order to integrate the information in the premises into a unified representation. Similar results are reported, for instance, in Carreiras & Santamaría (1997) and in an experiment from our own group. In our study, there was no significant difference in the percent of errors between continuous (39.7%) and semi-continuous (40.1%) premise orders, but both were significantly easier than the discontinuous order, which lead to 50.0% errors on average. Moreover, the data on premise processing times showed that the discontinuous premise order reliably increases the processing time for the third premise, because information from all premises must be integrated at this point (see Table 1, Exp. 1 from Knauff et al. 1998). Similar findings are reported from experiments in which the order of the terms within the premises was varied rather than the order of the premises. In parallel to the effect of premise order, these studies also indicate that the difficulty of reasoning tasks depends on the cognitive effort needed to integrate the premise information into a unified mental representation (Exp. 2 from Knauff et al. 1998).

The strongest argument in support of premise integration is the difference between determinate tasks, in which only a single model can be

Table 1
Premise processing times for the first, second, and third premises in the tasks with continuous, semi-continuous, and discontinuous premise order from Knauff et al. (1998)

Premise order	Premise 1	Premise 2	Premise 3
continuous	13.0	11.2	10.9
semi-continuous	13.6	11.0	14.4
discontinuous	12.4	13.9	19.5

constructed (as in our fMRI studies) and indeterminate tasks that call for multiple models. Byrne & Johnson-Laird (1989) compared such problems and found that indeterminate problems (34 % correct) are reliably harder than determinate problems (61 % correct). According to the mental model theory, indeterminate problems are more difficult because the construction of more than one integrated representation is more difficult than constructing a single model.

In our group, we have extensively investigated reasoning with indeterminate problems, and may have found the most convincing evidence for premise integration. The mental model theory ought to explain the integration process as a serial process that always produces the same first mental model. Hence, we tested the assumption of the existence of generally *preferred mental models* in an experiment, in which subjects had to determine possible relationships between objects based on the information given in the premises. The indeterminate problems called for three, five, or nine possible models. The results showed that whenever a reasoning problem has multiple solutions, reasoners prefer one of them and that individuals consistently prefer the same solution. This suggests that participants indeed integrate the information from the premises and inspect unified mental representations to find new information not given in the premises (Knauff et al. 1995, Rauh et al. 2005, Vandierendonck et al. 2004).

6. Explaining the visual-impedance effect

So far, we were only concerned with reasoning problems that invoke visual images. But what happens if the premises of a reasoning problem do not bias the reasoner to construct visual images? For example, they could straightforwardly lead to the spatial representations pertinent to reasoning without the phenomenal experience of an image. Are visual images *necessary* for reasoning? Do they have a *causal power* in the reasoning processes? Or are they only *epiphenomenona*, a side-effect of reasoning? The most convincing support for the three-stage theory is provided by a com-

bined behavioral and neuroimaging study that was specifically designed to answer these questions. In this study, we systematically investigated the engagement of mental imagery and the related brain areas during reasoning (Knauff et al. 2003). We speculated that only premises that are easy to visualize spontaneously elicit visual images, while other premises do not push reasoners to construct visual images. For instance, it is likely that reasoners construct a visual image from premises such as "The old lady with the yellow hut stands to the right of the man with the big ears" or even from "The Screwdriver is to the left of the hammer." But what about premises such as those in the second example from the introduction ("A is smarter than B", "B is smarter than C"). These premises are much more difficult to visualize and, therefore, probably no visual images are pressed into service during reasoning. Is reasoning easier or more difficult with these relations and does it activate different brain areas? In Knauff & Johnson-Laird (2002) we empirically identified four sorts of relations: (1) visuo-spatial relations that are easy to envisage visually and spatially, (2) visual relations that are easy to envisage visually but hard to envisage spatially, (3) spatial relations that are hard to envisage visually but easy to envisage spatially, and (4) control relations that are hard to envisage either visually or spatially. Then we started by conducting a series of behavioral experiments in which participants solved transitive inferences with these relations (Knauff & Johnson-Laird 2002). Apparently, the orthodox imagery theory would predict an advantage of visual and probably visuo-spatial relations. Our prediction, however, was that relations that elicit visual images containing details that are irrelevant to an inference should impede the process of reasoning, because the information pertinent to reasoning must be retrieved from the image. In contrast, relations that directly yield a spatial model without the "detour" of a visual image should speed up the process of reasoning in comparison with relations that elicit images. Our findings supported these predictions: In three experiments, we found that relations that are easy to visualize impaired reasoning. Reasoners were significantly slower with these relations than with the other sorts of relations. In fact, the spatial relations were the quickest, while the visual relations were the slowest. We called this the *visual-impedance effect* (Knauff & Johnson-Laird 2002). We then performed a brain imaging study using the same sorts of problems. As can be seen in Figure 2, all types of reasoning problems again evoked activity in the parietal cortices. This activity seems to be a "default mode" of brain functioning during reasoning, because individuals might have the facility to construct mental models from all sorts of relations. Such models will be spatial in form for visuospatial and spatial relations, and, as long-standing evidence suggests, even relations such as "smarter" are also likely to elicit spatial models (see, e.g. Johnson-Laird 1998, De Soto et al. 1965). However, only the problems

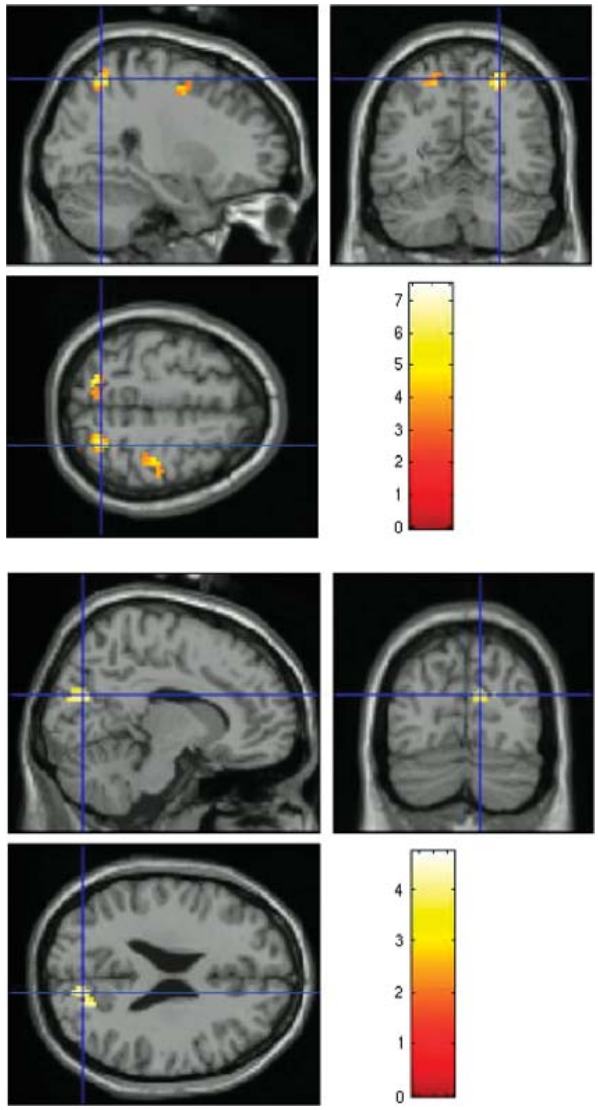


Fig. 2. Images representing differentially activated brain areas during reasoning. The three images above show the typical foci of activation resulting from reasoning with spatial relations. The location of the highlighted areas indicates that the spatial information from reasoning problems is mapped to areas of the brain responsible for the multimodal integration of space from perception and working memory. The three images below show the activity in the back of the brain suggesting that individuals naturally construct visual images, if the reasoning problem is easy to visualize (from: Knauff et al. 2003; see text for details).

based on visual relations also activated areas of the visual cortices. Presumably, in the case of visual relations, such as "The old lady with the yellow hat stands to the right of the man with the big ears" reasoners cannot suppress a spontaneous visual image of the appearance of the folks (certainly not members of the British royal family). Its construction calls for additional activity in visual cortices and retards the construction of a spatial mental model that is essential for the inferential process.

In a recent study with congenitally totally blind participants we collected remarkable extra evidence for this account. One consequence from the account is that people who are unable to construct visual images should not be disrupted by the visual details in the premises. We tested this hypothesis with a group of participants who were blind from birth. On the one hand, a visual account of reasoning might suggest that congenitally totally blind individuals—that do not experience visual mental images—should be impaired in reasoning with highly visual premises (e.g. Fraiberg 1980). On the other hand, there are several studies showing that persons who are blind from birth differ from sighted people in their use of *visual* images, but that they are as good as sighted in the construction of *spatial representations* (e.g. Kerr 1983). In particular, premises which are highly visual for sighted persons are unlikely to be visualized by persons who are blind from birth, and thus, we predicted, should not hinder their reasoning, because they are able to construct spatial representations without being sidetracked by irrelevant visual images. In Knauff & May (in press) we found exactly this difference between sighted and congenitally totally blind individuals. We tested a group of sighted participants, a group of congenitally totally blind participants, and a group of blindfolded participants with normal vision. For both, the sighted and blindfolded participants, the visual premises significantly impeded the process of reasoning in terms of both accuracy and time needed to verify the conclusion. The participants who were blind from birth, however, were not affected by the ease with which the verbal relations could be visualized. They showed the same reasoning performance across all types of problems. Obviously, people who are blind from birth are *immune* to the visual-impedance effects, since they do not tend to construct disrupting visual images from the premises.

7. Conclusions: Visual images can be a nuisance in reasoning

Psychological theories occasionally benefit when our introspective experiences agree with them. However, cognitive psychologists (and sometimes non-specialists) know very well that such a coincidence can be fatally misguiding. Moreover, people typically do not distinguish between differ-

ent types of introspection: representational states and cognitive operations (Barsalou 1999). The aim of this article thus was to clarify the role of visual and spatial representations in reasoning in an experimental fashion. As a starting point, in this paper only relational reasoning has been used and thus the presented theory is certainly limited to such inferences. Nevertheless, there is much evidence that other forms of reasoning also rely on mental models and that even more complex thinking succeeds without visual images although they are subjectively experienced. For instance, people often report representing mechanical systems and how they operate in visual mental images. However, Hegarty (2004) provides convincing evidence that mechanical reasoning—although it is frequently accompanied by imagery—is not a process of inspecting a holistic visual image in the "mind's eye." Instead, the "mental simulation" includes representations of non-visible properties and is even more efficient with non-imagery processes and spatial representations (Hegarty 2004). Given this converging evidence from different research areas, I can now envisage a research program that extends its attention to the role of spatial and visual representations in syllogistic reasoning (with quantifiers such as "all," "some," and "none"), modal reasoning (about what is possible and what is necessary), counterfactual reasoning (about hypothetical or imaginary cases), probabilistic reasoning (in which premises and conclusions have more than two truth values), temporal reasoning (about events that might have happened in the past or will happen in future), and inductive reasoning (in which a general rule is drawn from a large number of situations). In any case the current approach resolves many inconsistencies in the previous literature, because it shows that the visual characteristics of the premises can affect the process of inference. In agreement with the three-stage theory of reasoning suggested here, the reported studies demonstrate that reasoning is based on spatial representations, even if the content of premises elicit visual imagery that is not pertinent to reasoning. The spatial representations are represented and maintained in the parietal cortices. Here they are also inspected to find new information not given in the premises. These neural computations are performed under the regime of dorsolateral prefrontal brain areas. If, however, the spatial information must be retrieved from a visual image in order to construct the appropriate spatial mental model, additional processes come into play. The visual images activate occipito-temporal brain areas and the process of retrieving and integrating the spatial relations is realized by computations in AFC. These processes can be difficult because an image contains a large amount of visual details that is irrelevant for the reasoning process. Hence, it is likely that a visual image can even impede the process of reasoning. In contrast, if the content of the premises does not push reasoners towards constructing visual images, reasoning proceeds smoothly with spatial representations. One advantage of this approach is

that it is consistent with behavioral findings and data from neuroscientific research. In this way, it allows to formulate assumptions concerning the time course of reasoning processes and at the same time overcomes the naïve belief that neuroimaging data alone can explain how human cognitive processes work. One consequence from the approach is that individuals might not be aware of spatial representations during reasoning, or experience them as visual images, although they underlie our reasoning abilities. A second corollary is that visual imagery is *not* a mere epiphenomenon playing no causal role in reasoning (Pylyshyn 1981, 2002; see also: Knauff, Fangmeier, Ruff & Sloutsky 2005). It can be a nuisance because it impedes reasoning.

References

Andersen, R. A. (1997), 'Multimodal integration for the representation of space in the posterior parietal cortex', *Philosophical transactions of the Royal Society of London. Series B: Biological Sciences* **352**, 1421–1428.

Andersen, R., Snyder, L., Bradley, D. & Xing, J. (1997), 'Multimodal representation of space in the posterior parietal cortex and its use in planning movements', *Annual Review of Neuroscience* **20**, 303–30.

Antonietti, A. (1991), Why does mental visualization facilitate problem-solving?, *in* R. H. Logie & M. Denis, eds, 'Mental Images in Human Cognition', Elsevier Science Publishers, Amsterdam, pp. 211–227.

Barsalou, L. (1999), 'Perceptual symbol systems', *Behavioral and Brain Sciences* **22**, 577–660.

Block, N. (1981), *Imagery*, MIT Press, Cambridge, MA.

Brodmann, K. (1909), *Vergleichende Lokalisationslehre der Großhirnrinde*, J. A. Barth, Leipzig.

Bühler, K. (1909), 'Zur Kritik der Denkexperimente', *Zeitschrift für Psychologie* **51**, 108–118.

Burgess, N., Maguire, E. A., Spiers, H. J. & O'Keefe, J. (2001), 'A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events', *Neuroimage* **14**, 439–53.

Bushara, K. O., Weeks, R. A., Ishii, K., Catalan, M.-J., Tian, B., Rauschecker, J. P. & Hallett, M. (1999), 'Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans', *Nature Neuroscience* **2**, 759–766.

Byrne, R. M. J. & Johnson-Laird, P. N. (1989), 'Spatial reasoning', *Journal of Memory and Language* **28**, 564–575.

Carreiras, M. & Santamaría, C. (1997), 'Reasoning about relations: spatial and nonspatial problems', *Thinking and Reasoning* **3**, 191–208.

Chen, W., Kato, T., Zhu, X.-H., Ogawa, S., Tank, D. W. & Ugurbil, K. (1998), 'Human primary visual cortex and lateral geniculate nucleus activation during visual imagery', *NeuroReport* **9**, 3669–3674.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J. & Gabrieli, J. D. E. (2001), 'Rostrolateral prefrontal cortex involvement in relational integration during reasoning', *NeuroImage* **14**, 1136–49.

Clement, C. A. & Falmagne, R. J. (1986), 'Logical reasoning, world knowledge, and mental imagery: Interconnections in cognitive processes', *Memory & Cognition* **14**, 299–307.

Colby, C. L. & Duhamel, J. R. (1996), 'Spatial representations for action in parietal cortex', *Cognitive Brain Research* **5**(1-2), 105–115.

De Soto, L. B., London, M. & Handel, M. S. (1965), 'Social reasoning and spatial paralogic', *Journal of Personality and Social Psychology* **2**, 513–521.

Denis, M., Logie, R., Cornoldo, C., de Vega, M. & Engelkamp, H. (2001), *Imagery, Language and Visuo-Spatial Thinking*, Psychology Press, Hove (UK).

Ehrlich, K. & Johnson-Laird, P. N. (1982), 'Spatial descriptions and referential continuity', *Journal of Verbal Learning and Verbal Behavior* **21**, 296–306.

Evans, J. S. B. T., Newstead, S. E. & Byrne, R. M. J. (1993), *Human Reasoning: The Psychology of Deduction*, Lawrence Erlbaum Associates, Hove (UK).

Fangmeier, T., Knauff, M., Ruff, C. C. & Sloutsky, V. (in press), 'fMRI evidence for a three-stage model of deductive reasoning', *Journal of Cognitive Neuroscience* .

Farah, M. J., Hammond, K. M., Levine, D. N. & Calvanio, R. (1988), 'Visual and spatial mental imagery: Dissociable systems of representation', *Cognitive Psychology* **20**, 439–462.

Fraiberg, S. H. (1980), *Insights to the Unseen World*, Basic Books, New York.

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

Goel, V., Büchel, C., Frith, C. & Dolan, R. (2000), 'Dissociation of mechanisms underlying syllogistic reasoning', *NeuroImage* **12**, 504–514.

Goel, V. & Dolan, R. J. (2001), 'Functional neuroanatomy of three-term relational reasoning', *Neuropsychologia* **39**, 901–909.

Goel, V., Gold, B., Kapur, S. & Houle, S. (1997), 'The seats of reason? An imaging study of deductive and inductive reasoning', *NeuroReport* **8**, 1305–1310.

Goel, V., Gold, B., Kapur, S. & Houle, S. (1998), 'Neuroanatomical correlates of human reasoning', *Journal of Cognitive Neuroscience* **10**, 293–302.

Halford, G. S., Wilson, W. H. & Phillips, S. (1998), 'Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology', *Behavioral and Brain Sciences* **21**, 803–831.

Hegarty, M. (2004), 'Mechanical reasoning by mental simulation', *Trends in Cognitive Science* **8**(6), 280–285.

Huttenlocher, J. (1968), 'Constructing spatial images: A strategy in reasoning', *Psychological Review* **75**, 550–560.

Johnson-Laird, P. (2001), 'Mental models and deduction', *Trends in Cognitive Science* **5**, 434–442.

Johnson-Laird, P. N. (1972), 'The three-term series problem', *Cognition* **1**, 57–82.

Johnson-Laird, P. N. (1983), *Mental Models*, Cambridge University Press, Cambridge.

Johnson-Laird, P. N. (1998), Imagery, visualization, and thinking, *in* J. Hochberg,

ed., 'Perception and Cognition at Century's End', Academic Press, San Diego, CA, pp. 441–467.

Johnson-Laird, P. N. & Byrne, R. (1991), *Deduction*, Erlbaum, Hove (UK).

Johnson-Laird, P. N., Byrne, R. & Tabossi, P. (1989), 'Reasoning by model: The case of multiple quantifiers', *Psychological Review* **96**, 658–673.

Kerr, N. H. (1983), 'The role of vision in "visual imagery" experiments: Evidence from the congenitally blind', *Journal of Experimental Psychology: General* **112**, 265–277.

Knauff, M. (1999), 'The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning', *Spatial Cognition and Computation* **1**, 261–290.

Knauff, M., Fangmeier, T., Ruff, C. C. & Johnson-Laird, P. N. (2003), 'Reasoning, models, and images: Behavioral measures and cortical activity', *Journal of Cognitive Neuroscience* **4**, 559–573.

Knauff, M., Fangmeier, T., Ruff, C. & Sloutsky, V. (2005), fMRI evidence for a three-stage model of deductive reasoning, *in* 'Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society', Erlbaum, Mahwah, NJ.

Knauff, M., Jahn, G. & Vosgerau, G. (2005), Indeterminacy in spatial reasoning and working memory. Manuscript submitted for publication.

Knauff, M. & Johnson-Laird, P. N. (2002), 'Visual imagery can impede reasoning', *Memory & Cognition* **30**, 363–371.

Knauff, M., Kassubek, J., Mulack, T. & Greenlee, M. W. (2000), 'Cortical activation evoked by visual mental imagery as measured by functional MRI', *NeuroReport* **11**, 3957–3962.

Knauff, M. & May, E. (in press), 'Mental imagery, reasoning, and blindness', *Quarterly Journal of Experimental Psychology* .

Knauff, M., Mulack, T., Kassubek, J., Salih, H. R. & Greenlee, M. W. (2002), 'Spatial imagery in deductive reasoning: A functional MRI study', *Cognitive Brain Research* **13**, 203–212.

Knauff, M., Rauh, R. & Schlieder, C. (1995), Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allen's calculus, *in* 'Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society', Erlbaum, Mahwah, NJ, pp. 200–205.

Knauff, M., Rauh, R., Schlieder, C. & Strube, G. (1998), Mental models in spatial reasoning, *in* C. Freska, C. Habel & K. F. Wender, eds, 'Spatial Cognition - An Interdisciplinary Approach to Representing and Processing Spatial Knowledge', Lecture Notes in Computer Science, Vol. 1404, Subseries: Lecture Notes in Artificial Intelligence, Springer, Berlin, pp. 267–291.

Kosslyn, S. M. (1980), *Image and Mind*, Harvard University Press, Cambridge, MA.

Kosslyn, S. M. (1994), *Image and Brain*, MIT Press, Cambridge, MA.

Kosslyn, S. M., Alpert, N. M., Thompson, W. L., Maljkovic, V., Weise, S. B., Chabris, C. F., Hamilton, S. E., Rauch, S. L. & Buonanno, F. S. (1993), 'Visual mental imagery activates topographically organized visual cortex: PET investigations', *Journal of Cognitive Neuroscience* **5**, 263–287.

Kosslyn, S. M., Chabris, C. F., Marsolek, C. J. & Koenig, O. (1992), 'Categorical versus coordinate spatial relations: Computational analyses and computer simulations', *Journal of Experimental Psychology: Human Perception and Performance* **18**, 562–577.

Kosslyn, S. M., Koenig, O., Barrett, A., Backer Cave, C., Tang, J. & Gabrieli, J. D. E. (1989), 'Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations', *Journal of Experimental Psychology: Human Perception and Performance* **15**, 723–735.

Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Thompson, W. L., Ganis, G., Sukel, K. E. & Alpert, N. M. (1999), 'The role of area 17 in visual imagery: Convergent evidence from PET and rTMS', *Science* **284**, 167–170.

Kosslyn, S. M., Thompson, W. L. & Alpert, N. M. (1997), 'Neural systems shared by visual imagery and visual perception: A positron emission tomography study', *NeuroImage* **6**, 320–334.

Manktelow, K. (1999), *Reasoning and Thinking*, Psychology Press, Hove (UK).

Mellet, E., Petit, L., Mazoyer, B., Denis, M. & Tzourio, N. (1998), 'Reopening the mental imagery debate: Lessons from functional anatomy', *NeuroImage* **8**, 129–139.

Newstead, S. E., Pollard, P. & Griggs, R. A. (1986), 'Response bias in relational reasoning', *Bulletin of the Psychonomic Society* **2**, 95–98.

Paivio, A. (1971), *Imagery and Verbal Processes*, Holt, Rinehart, and Winston, New York [Reprinted (1979), Erlbaum, Hillsdale, NJ].

Paivio, A. (1986), *Mental Representations: A Dual Coding Approach*, Oxford University Press, New York.

Perky, C. W. (1910), 'An experimental study of imagination', *Journal of Psychology* **21**, 422–452.

Petrides, M. (2000), 'The role of the mid-dorsolateral prefrontal cortex in working memory', *Experimental Brain Research* **133**, 44–54.

Pylyshyn, Z. W. (1981), 'The imagery debate: Analogue media versus tacit knowledge', *Psychological Review* **88**, 16–45.

Ragni, M., Knauff, M. & Nebel, B. (2005), A computational model of human reasoning with spatial relations, *in* 'Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Scociety', Erlbaum, Mahwah, NJ.

Rauh, R., Hagen, C., Knauff, M., Kuß, T., Schlieder, C. & Strube, G. (2005), 'Preferred and alternative mental models in spatial reasoning', *Spatial Cognition and Computation* **in press**.

Ruff, C. C., Knauff, M., Fangmeier, T. & Spreer, J. (2003), 'Reasoning and working memory: Common and distinct neuronal processes', *Neuropsychologia* **41**, 1241–1253.

Sabbah, P., Simond, G., Levrier, O., Habib, M., Trabaud, V., Murayama, N., Mazoyer, B. M., Briant, J. F., Raybaud, C. & Salamon, G. (1995), 'Functional magnetic resonance imaging at 1.5 T during sensorimotor and cognitive tasks', *European Journal of Neurology* **35**, 131–136.

Sadoski, M. (1985), 'The natural use of imagery in story comprehension and recall: Replication and extension', *Reading Research Quarterly* **20**, 658–667.

Sadoski, M. & Paivio, A. (1994), A dual coding view of imagery and verbal processes in reading comprehension, *in* R. B. Ruddell, M. R. Ruddell & H. Singer, eds, 'Theoretical Models and Processes of Reading', 4th edn, International Reading Association, Newark, DE, pp. 582–601.

Schlieder, C. (1999), The construction of preferred mental models in reasoning with interval relations, *in* G. Rickheit & C. Habel, eds, 'Mental Models in Discourse Processing and Reasoning', Elsevier, Amsterdam, pp. 333–357.

Schlieder, C. & Berendt, B. (1998), Mental model construction in spatial reasoning: A comparison of two computational theories, *in* U. Schmid, J. F. Krems & F. Wysotzki, eds, 'Mind Modelling: A Cognitive Science Approach to Reasoning, Learning, and Discovery', Pabst Science Publishers, Lengerich, pp. 133–162.

Shaver, P., Pierson, L. & Lang, S. (1975), 'Converging evidence for the functional significance of imagery in problem solving', *Cognition* **3**, 359–375.

Shepard, R. N. & Cooper, L. A. (1982), *Mental Images and Their Transformations*, MIT Press, Cambridge, MA.

Stanfield, R. A. & Zwaan, R. A. (2001), 'The effect of implied orientation derived from verbal context on picture recognition', *Psychological Science* **12**, 153–156.

Sternberg, R. J. (1980), 'Representation and process in linear syllogistic reasoning', *Journal of Experimental Psychology: General* **109**, 119–159.

Suler, J. R. & Riziello, J. (1987), 'Imagery and verbal processes in creativity', *Journal of Creative Behaviour* **21**, 1–6.

Titchener, E. B. (1909), *Lectures on the Experimental Psychology of the Thought Processes*, Macmillan, New York.

Tye, M. (1991), *The Imagery Debate*, MIT Press, Cambridge, MA.

Ungerleider, L. & Mishkin, M. (1982), Two cortical visual systems, *in* D. J. Ingle, M. A. Goodale & R. J. W. Mansfield, eds, 'Analysis of Visual Behaviour', MIT Press, Cambridge, MA, pp. 549–586.

Vandierendonck, A., Dierckx, V. & De Vooght, G. (2004), 'Mental model construction in linear reasoning: Evidence for the construction of initial annotated models', *Quarterly Journal of Experimental Psychology* **57A**, 1369–1391.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., Thomas, C. R. & Miller, B. L. (1999), 'A system for relational reasoning in human prefrontal cortex', *Psychological Science* **10**, 119–25.

Watson, J. B. (1913), 'Psychology as the behaviorist views it', *Psychological Review* **20**, 158–177.

Xing, J. & Andersen, R. A. (2000), 'Models of the posterior parietal cortex which perform multimodal integration and represent space in several coordinate frames', *Journal of Cognitive Neuroscience* **12**, 601–614.

Zwaan, R. A., Stanfield, R. A. & Yaxley, R. H. (2002), 'Do language comprehenders routinely represent the shapes of objects?', *Psychological Science* **13**, 168–171.

Part III

Perception, Emotion, and Language

This Page is Intentionally Left Blank

Carsten Held, Markus Knauff, Gottfried Vosgerau

Introduction: Perception, Emotion, and Language

This part of the book describes some novel perspectives on the connection between mental models and perception, emotion, and language. The main question is whether the general picture outlined in the first parts of the book also helps to understand mental phenomena other than reasoning and learning. The answers are quite different: Mental models seem to provide a very promising framework for emotions, but the case of visual perception calls for further refinements and additional theorizing. Furthermore, syntactic and semantic processing of natural language challenges some of the basic ideas of mental model theory.

The first contribution investigates the place of mental models in visual perception. We perceive the space around us as Euclidean and three-dimensional. In order to do so, several depth cues are evaluated to transform the two-dimensional retinal picture into a three-dimensional representation. **Rehkämper** discusses the case of visual picture perception, where the perceived object is already two-dimensional. According to Rehkämper, if mental models are analogical representations of the space around us, three-dimensional models in perception should be Euclidean as well. However, empirical evidence makes questionable whether the perceived space has a geometry at all. At least, it seems quite certain that it is not Euclidean. Hence, if mental models play a role in perception at all, it has to be explained how a Euclidean model can be constructed on the basis of non-Euclidean representations. One possibility is the technique of modified weak fusion, which would have to operate at an unconscious level. But even if there is such a mechanism, there are differences between perceptual models and models used in reasoning: the former are transparent and cannot be modified, the latter are opaque and modifying them is essential for reasoning. Therefore, a unified account of mental models in perception

and reasoning faces several serious problems, the solution of which requires much further refinement of the theory.

The interdependence of emotion and cognition is becoming more and more evident from recent empirical research in different areas. Yet, the characterization of emotions in terms of a representational theory of mind does not seem to be satisfactory. **Pauen** develops three constraints for such a representational theory of emotions, based on a review of recent empirical findings. First, similar emotions are evoked in similar situations, which fact suggests a "similar input - similar output" principle. Second, representations of emotions have to be dynamic, since not only situations but primarily (outcomes of) processes are emotionally evaluated. Third, emotions are multi-modal, i.e. one emotion can be triggered by inputs of different modalities. Mental models meet all of the three constraints. Hence, they are much more suitable for a representational theory of emotions than symbolic theories. Moreover, mental model theory is developed for the realm of cognition, such that the interaction of cognition and emotion could be easily described if the underlying representations were of the same kind. Pauen concludes that mental models are the most promising framework for a theory of the interdependence of emotion and cognition. Further empirical research will provide details to fill in the rather metaphorical character of the theory so far.

If we think about the world in terms of mental models it seems reasonable to assume that syntactic and semantic processing of natural language is constituted by the construction of such mental models. **Hemforth and Konieczny** show that such an eliminative view (that denies that other mechanisms and representations are involved) must leave unexplained a large amount of data. In particular, preferences for certain interpretations of ambiguous sentences are dependent on different syntactical and contextual factors as well as on background knowledge. The parsimony principle for mental models cannot explain the diversity of effects. Moreover, the representation of numbers is hardly captured by mental models: It is unclear how vague quantifiers (e.g. "quite a lot") and exact large numbers could be represented within mental models. Additionally, anaphora resolution in texts seems to be dependent on several semantic as well as syntactic levels of representation. Assuming only mental models does not suffice. All in all, mental models cannot be the (exclusive) basis for syntactic and semantic processing. Different levels of processing have to be assumed. Moreover, in order to determine the place of mental models in this hierarchy of representations, much more detailed constraints have to be added to the general picture.

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Pictures, Perception, and Mental Models

Klaus Rehkämper [1]

Carl-von-Ossietzky-Universität Oldenburg [2]

Abstract
The main point of this paper is to consider the way space is perceived in pictures and in "reality" and the question of whether mental models are a good means in explaining how space is visually perceived. Real or physical space is presumed to be (locally) Euclidean. Some kinds of pictures—e.g. pictures in perspective—are lawfully connected to the depicted scene so that the (Euclidean) geometry of that scene is preserved in these pictures. Following Johnson-Laird, visual perception is based on the construction of a partially analogical mental model. Therefore, as I will show, the geometry of a mental model representing the spatial layout of a scene in the physical world (or of a picture of such a scene) should also be Euclidean. However, at least since the famous experiments of Blumenfeld in 1913 it seems clear that our phenomenal or visual space is not Euclidean. How does this fit together? Can it be that the different cues which are involved in the perception of spatial arrangements are not modeled in a Euclidean way, but that the model in toto is (nearly) Euclidean? Is such a model built up by using "modified weak fusion"?

[1] I would like to thank Verena Gottschling, Carsten Held, Wolfgang Huemer, and Markus Knauff for helpful comments.
[2] E-mail: klaus.rehkaemper@uni-oldenburg.de

1. Introduction

The thoughts presented here are based on three common assumptions:[3]

(1) Visual perception is inverted optics.
(2) A theory of visual perception must include a theory of picture perception.[4]
(3) Physical space is (locally) Euclidean.

According to the first claim, visual perception involves a process of reconstructing a three-dimensional scene from purely two-dimensional patterns of light rays. The second assumption states that a valid theory of visual perception has to explain how information is extracted from perceiving not only reality, but from a pictorial representation of it as well. The third claims that the geometry for describing the structure of the space surrounding us (i.e. we are not talking about astrophysics, the universe and things like that) is a Euclidean one. This means especially that parallel lines always have the same distance and intersect only in infinity.

It seems tempting to hold these natural presupposition against a theory for explaining visual perception that has already shown its advantages for explaining other cognitive abilities as text understanding or the carrying-out of logical inferences—the theory of mental models as introduced by Johnson-Laird (1983). Can mental models be used to explain how visual perception works? Johnson-Laird at any rate holds this view.

My argument will proceed in three steps. First, I will shortly describe the problem of stereopsis and the different cues used in perceiving depth and spatial layouts. I will argue that there are nine different cues for perceiving depth, five of which can also be used to recognize depth in pictures. Second, I will concentrate on some problems of geometry. What does it mean, that physical space is Euclidean and phenomenal space is perhaps not? What alternatives are there? In the final part, I will scrutinize the idea that mental models are a good means to explain the process of visual perception.

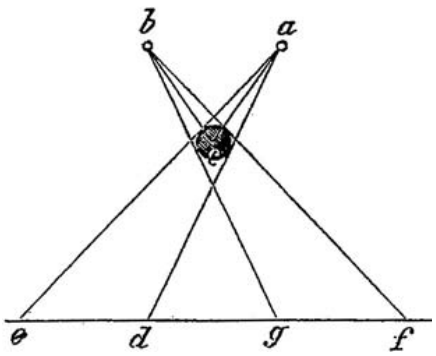
---

[3] The question raised in this paper was partly inspired by Cutting & Vishton (1995, 70): "[H]ow do we come to perceive the three-dimensional layout of our environment with reasonable, even near metric, accuracy when taken singly, none of the visual sources of information yields metric information throughout the range of distances we need?"

[4] Although this assumption is sometimes questioned (Rogers 1995, cf.), it is accepted by the vast majority of psychological theories of picture perception, even if they are starkly contrasting otherwise.

2. The problem of stereopsis

Before I start my examination, I will prepare the ground for my argument. How do pictures represent the physical world?[5] The first thing to mention is that pictures are usually two-dimensional representations of three-dimensional scenes and therefore can only show what we see in monocular vision. Leonardo da Vinci was the first to note this fact:[6]



> It is impossible that a picture copying outlines, shade, light and colour with the highest perfection can appear to posses the same relief as that which appears in the object in nature, unless this natural object is looked at over the long distance and with a single eye. This is proven as follows: let the eyes be *a* and *b*, looking at an object *c*, with the converging central axes of the eyes as *ac* and *bc*, which converge on the object at the point *o*. The other axes, lateral to the central one, see the space *gd* behind the object, and the eye *a* sees all the space *fd*, and the eye *b* sees all the space *ge*. Hence the two eyes see behind the object and all the space *fe*. [...] This cannot happen with someone who looks at an object with one eye. [...] [S]omething painted interrupts our view of all the space behind it, and in no way is it possible to see any part of the background behind it. (Leonardo CU 155v-156r, Kemp 2001, 63f.)

We do not perceive—binocularly—picture space in the same way we perceive physical space, which is located directly before our eyes. An "object in nature" close to us appears to be transparent, whereas objects in pictures

[5] I will restrict my inquiry to pictures in perspective. There is no doubt that if pictures are able to represent space faithfully, these pictures do.

[6] Nearly all the people writing about the problem of stereopsis refer back to Wheatstone (e.g. 1838), who himself uses this very thought of Leonardo as his starting point.

are always opaque. This means, that in the perception of physical objects we can perceive space and objects, which are situated behind a particular object, whereas in a picture objects in the foreground will occlude other objects. [7]

Looking with two eyes gives us two slightly different retinal images. [8] These different images give rise to a full-fledged three-dimensional impression of the scene perceived. This phenomenon is nowadays called 'binocular disparity' and explains, among other things, how a stereogram works. Each eye is stimulated by a slightly different pattern of light rays, which leads to slightly different retinal images.

Leonardo da Vinci was also the first to describe the three different kinds of perspective that can be used for producing a picture that faithfully represents space.

> Perspective is divided into three parts, of which the first is concerned solely with the outlines of the bodies; the second in the diminution of colours at varying distances; the third in the loss of definition of bodies at various distances.
>
> Now, the first, which only embraces the outlines and contours of bodies, is called drawing, that is to say, the figuration of any solid body. From this arises another science, which embraces light and shade, or we may wish to say *chiaroscuro*, a science of complex position. (Leonardo CU 2v-r, Kemp 2001, 16)

Today the first kind of perspective is called central or linear perspective and belongs completely to the realm of geometry. Pictures that involve this kind of perspective usually have a central vanishing point and the representations of parallel lines converge in the picture. [9] Alberti, Piero della Francesca and Leonardo, to name just a few, developed (linear) perspective as a formal technique of realistic painting during the renaissance.

The two other kinds of perspective are subsumed today under the name "aerial perspective". A wonderful example for the use of this technique can

[7] Try it yourself. Just stretch out one arm, your thumb should be up. Now look—using only one eye—at your thumb; the space behind is occluded. Now use both eyes. All the space behind your thumb becomes visible; the finger itself appears to be transparent.

[8] In talking of retinal images I follow the common way of speaking, but let me put straight the fact that retinal images are not pictures; they are (mirror) reflections. But the theory of central projection explains the structure of pictures in perspective as well as the structure of the patterns of light rays, which reach the retina and are partially reflected by it. To be precise, the retinal image we see, when we look in the eye of someone using a suitable instrument, is exactly the amount of light, which is **not** used in the process of visual perception. We do not see our retinal images and they play no part in the process of visual perception. (For a more detailed description of the role of perspective see e.g. Rehkämper 2003a, 2003b.)

[9] If they are not parallel to the picture plane as well.

Fig. 1. Caspar David Friedrich "Morgen im Riesengebirge" 1810-11 (orig. in color) Alte Nationalgalerie, Berlin

be found in Caspar David Friedrich's painting "Morgen im Riesengebirge" (Fig. 1). We notice that with increasing distance the color of the mountains gets lighter and the edges lose contour. [10] In comparison to Uccello's "The Rout of San Romano" (Fig. 2) the usefulness of this technique becomes even more prominent. The background in Uccello's picture is not painted using aerial perspective. As a result, it does not give the same convincing impression of depth.

Perceiving a picture of a scene can never be the same as perceiving the real scene directly, because a picture only presents a view that is equivalent (*cum grano salis*) to monocular vision. Looking at the world around us in the 'usual way', we perceive things binocularly. We do this via two retinal 'images,' which are in a way similar to two slightly different pictures in perspective.

There are several techniques to improve the impression of depth in a picture, all of which have their equivalents in the visual perception of depth. In modern introductory books of the psychology of visual perception (e.g. Goldstein 1996), usually nine possible cues for perceiving depth are described:

(1) Occlusion
(2) Relative size
(3) Relative density
(6) Binocular disparity
(7) Motion parallax
(8) Convergence

[10] If the picture were in color, we would also notice that the colors becomes more bluish.

Fig. 2. Paolo Uccello "The Rout of San Romano" 1456 (orig. in color) National Gallery, London

(4) Height in the visual field (9) Accommodation
(5) Aerial perspective

Linear perspective can be accounted for as the combination of (1), (2), (3), and (4) plus converging lines (plus perhaps (5) aerial perspective).

Cues (1)–(5) could be used for depth perception in pictures as well as for perceiving depth in physical space; i.e. picture perception and perceiving 'the real thing' differs only in respect to cues (6)–(9).

But relative to the distance of the objects perceived these cues are neither equally valid nor does the value of information stay constant. As James Cutting observes: "Different sources of information seem to work differently" (Cutting 2003, 223, cf. Table 1). Accommodation and convergence, for example, are useful, when the object is very close, whereas aerial perspective becomes so only in cases where the object is 100m or more away. Furthermore, Cutting (e.g. 2003) differentiates between three regions in which the effectiveness of the different depth cues differs considerably. Following Cutting, the egocentric (physical) space should roughly be divided into

(I) a personal space (which has a radius of approx. 2m),
(II) an action space (with a radius of approx. 30m), and
(III) a vista space (with a radius > 30m).

No matter how far away the objects are, occlusion, relative size, and density seem to be equally efficacious; in personal space we additionally make use of binocular disparity, motion perspective, accommodation, and convergence; in action space height in the visual field and motion perspective

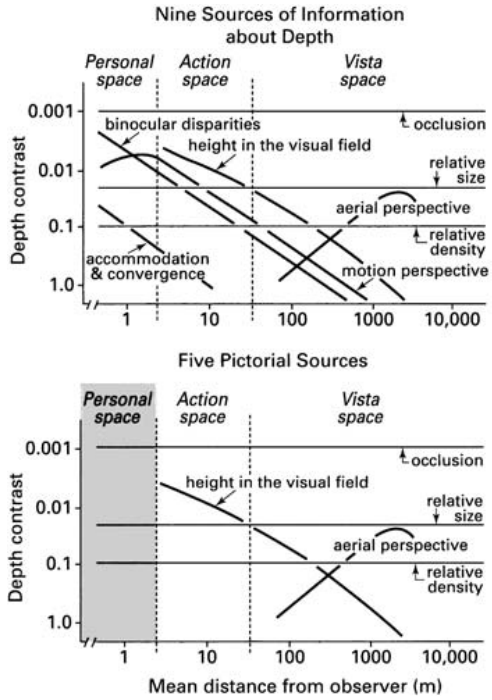


Table 1
Threshold functions for pairwise ordinal distance judgments are shown for nine sources of information. The data are plotted as a function of the mean distance of two objects from the observer (log transformed) and of their depth contrast ($(d_1 - d_2) / [(d_1 + d_2) / 2]$). In the lower panel personal space is omitted, because only few pictures or paintings show objects in this region. (Cutting 2003, 223)

are more helpful; and in vista space aerial perspective becomes—apart from occlusion, relative size and density—the most valuable cue (Table 1). And, as Cutting also points out, none of this information has to be metric; an ordinal ordering is all we need (Cutting 2003, 236). And this is exactly what a picture in perspective offers to an observer.

As an intermediary result we can conclude that perceiving depth (and spatial layouts) in pictures and perceiving depth (and spatial layouts) in reality are closely related, or as Cutting (2003, 236) puts it: "[They] are cut from the same informational cloth." There are surely differences, but they are outweighed by the accordances, especially in the case of perspectival pictures (cf. Hecht et al. 1999, Rogers 2003).

3. A question of geometry

3.1. The geometry of physical space

What does it mean to ask whether space is structured in a Euclidean or non-Euclidean way? Let us begin with the famous five axioms of Euclid's "Geometry":[11]

(A-1) Every two points lie on exactly one line.

(A-2) Any line segment with given endpoints may be continued in either direction.

(A-3) It is possible to construct a circle with any point as its centre and with a radius of any length.

(A-4) If two lines cross such that a pair of adjacent angles are congruent, then each of these angles are also congruent to any other angle formed in the same way.

(A-5) (Parallel Axiom): Given a line l and a point not on l, there is one and only one line, which contains the point, and is parallel to l.

Since school days, we all are more or less familiar with these axioms. They appear to be unproblematic. But right from Euclid's days on, the Parallel Axiom was under discussion. It did not seem to be self-evident in the way the others are, and in the following two and a half millennia many attempts were made to show that it can be deduced from the other four.[12]

However, it was not until the middle of the 19th century that three mathematicians—Gauß, Lobachevsky and Riemann—could show independently from each other that it is impossible to deduce Axiom 5 from the others and moreover that this axiom can be replaced by two alternatives.

Gauß and Lobachevsky replaced Axiom 5 by:

(A-5H) (Hyperbolic Geometry Parallel axiom): Given a line l and a point not on l, there exist at least two distinct lines that contain the point and are parallel to l.

For a given line there are at least two (in fact infinitely many) lines that do not intersect the given line at some point.

The famous German mathematician Riemann presented a different alternative:

(A-5S) (Spherical Geometry Parallel axiom): Given a line l and a point not on l, there exists no line that contains the point and is parallel to l.

In a world where such a geometry is valid, no line exists that does not intersect l and therefore is parallel to it.

[11] The following is not the original, but a modern but formal equivalent formulation.

[12] For a brief history of non-Euclidean geometry see Trudeau (1985).

Until this time it was taken for granted that Euclidean geometry was the only geometry possible. [13] But after the findings of Gauß, Lobachevsky, and Riemann, and especially after Einstein's publication of the theory of relativity, which made use of non-Euclidean geometry, things looked different. This raises the question of what kind of geometry was the correct one to describe the structure of the physical world. As I said in the introduction, it is commonly accepted that as long as we talk about earthly affairs, Euclidean geometry seems to be the best choice.

But what about our phenomenal world? Do we perceive our environment as structured according to the axioms of Euclid?

3.2. Is binocular space Euclidean?

At least since the beginning of the 20th century, it is a well-known phenomenon that curved lines under specific conditions may actually appear straight to an observer. (e.g. Hillebrand 1902).

In 1913, the psychologist Walter Blumenfeld examined this phenomenon more closely. In two different tasks, subjects were asked to arrange very small lights in an otherwise darkened room. At first, in the 'parallel alley experiment,' they were assigned the task to arrange two receding rows of lights—presented in front of them at eye-level—in a way that the rows seemed to be parallel. In another set-up, the 'equidistance alley experiment,' the task was to arrange pairs of light—again presented in front of them at eye-level—in a way that the distance between each pair of lights was held constant. Whereas in the first task all lights were always visible all the time, in the second only two pairs of lights were visible at a time, with the first pair serving as reference for the others, which were presented one by one, each a bit farther away from the observer.

Since the distance between the reference lights was the same in both tasks, the outcome should have been two identical rows of lights. At least, this is what one would have expected, if the Euclidean geometry is the correct choice for describing the structure of our phenomenal space. The two rows were not identical—in fact the rows of the parallel alleys lay inside of the rows of the equidistance alleys, which leads to the assumption that the geometry of the visual space is hyperbolic—nor were the rows straight.

Over the years a long series of experiments confirmed Blumenfeld's results and the mathematical interpretations of the experiments led to the conclusion that the geometry of the phenomenal (or visual) space is hyperbolic with constant Gaussian curvature. Although—as the overview (Ta-

[13] That was one reason why, for example, Kant thought that the sentences of geometry are synthetic a priori. They tell us something about the world without being based on empirical knowledge.

Table 2
The Geometry of the visual space (based upon Suppes 1997, edited and extended by K.R.)

Name	Claim
Eucild (300 BC)	Theory of perspective
Reid (2000), Daniels (1972), Angell (1974)	Geometry of visibles is spherical
Kant (1998)	A priori Euclidean
Blumenfeld (1913)	Parallel alleys not identical to equidistance alleys
Luneburg (1947, 1950)	Visual space is hyperbolic
Gibson (1950)	Visual space is Euclidean
Blank(1957, 1961)	Ess. same as Luneburg
Hardy et al. (1953)	Ess. same as Luneburg
Schelling (1956)	Hyperbolic rel. to given fixation point
Gogel (1963)	Evidence for contextual geometry
Foley (1972)	Visual space is non-honogeneous
Indow (1967, 1974)	MDS-Method yield good Euclidean fit
Grünbaum (1963)	Questions Luneburg theory
Strawson (1996)	Phenomenal geometry is Euclidean
Wagner (1985)	Visual space is affine Euclidean
Zimmer (1998*a*, 1998*b*)	Visual space has perhaps no geometry

ble 2) shows—the interpretations of the data are not unanimous. Today, it seems to be more or less accepted that the geometry of the phenomenal space is not Euclidean. Perhaps the most astonishing interpretation of recently collected data is Zimmer's finding (Zimmer 1998*a*,*b*) that phenomenal space might be neither Euclidean nor hyperbolical or spherical. Zimmer tested the betweenness-structure (as described by Suppes et al. 1989), which makes use of only the first four of Euclid's axioms. It is thus a simple or naïve geometry, upon which both Euclidean and hyperbolic geometry rest. Consequently, if one can show that phenomenal space cannot be described by this simple geometry, the question of whether it is Euclidean or hyperbolic does not even arise.

Zimmer in fact found that not even the betweennes-structure seems to be applicable to phenomenal space. All in all "[t]he results show that, contrary to previous assumptions, neither Euclidean nor hyperbolic geometry serve as a valid representation for the whole extent of binocular space" (Zimmer

1998*b*, 393). But she concludes: "If the present outcome were attributable to the extremely reduced experimental situation, namely presenting three point-like sources of light in a completely dark surround, the introduction of a stable visual context should improve the structure of binocular space." (Zimmer 1998*b*, 398)

While physical space is Euclidean, phenomenal space is and is not at the same time. It is in fact true that under ecologically valid conditions, i.e. in situations in which nearly all possible cues for depth can be exploited, visual space seems to have a stable structure, which is nearly Euclidean. In situations where only one or two cues were tested, these single cues apparently do not obey the rules of a geometry—Euclidean, hyperbolic or whatsoever. But in an ecologically valid situation the 'teamwork' of the cues leads to a stable picture of the world. Moreover, the information provided by the single cues only has to be ordinal information (cf. Cutting 2003). So, how should the various cues for perceiving depth be combined in order to get a coherent representation of the world?

4. The geometry of models

In 1995, Landy et al. proposed their theory of *modified weak fusion*, which is based on a Bayesian approach (i.e. a probabilistic approach) to model the integration of the several depth cues (Landy et al. 1995). Included in this theory is a simple, linear cue combination rule (weighted averaging of depth estimation).

This "weak" fusion is further modified (and hence results in apparently nonlinear behavior) by three additional processes: (1) cue weights change from scene to scene in response to perceived changes in cue availability and reliability; (2) information from different cues is often incommensurate and cues are "promoted" to be on a comparable scale to be averageable; (3) cue weights may also change based on the estimates of depth themselves to achieve a cue combination rule robust against gross errors derived from individual cues (Landy & Maloney 1998, Landy et al. 1995). This fits nicely with the findings of Cutting (cf. 2003).

Let me now come back to the question raised at the beginning of this inquiry. Are mental models a good means to explain visual perception? Are all these findings coherent with what Johnson-Laird states about the properties of mental models?

What is a mental model? Essentially it is "an internal representation of a state of affairs in the external world" (Johnson-Laird 1992, 932). And in respect to visual perception Johnson-Laird adds:

> Mental models can be constructed on the basis of visual perception (Marr, 1982) ... Their essential characteristics are that *their structure corresponds to the structure of what they represent.* Like a diagram (Maxwell, 1911) or an architect's model, the parts of the model correspond to the relevant parts of what it represents, and *the structural relations between the parts of the model are analogous to the structural relations in the world.* Hence, a model represents a set of individuals by a set of mental tokens, it represents the properties of the individuals by the properties of the tokens, and it represents the relations among the individuals by the relations among the tokens. (Johnson-Laird 1998, 447, italics by K.R.)

> The representation must therefore be a *model* in three dimensions of the scene, which, like an architect's model, makes explicit the shape of everything in the scene. (Johnson-Laird 1988, 107)

The physical world is Euclidean and between the structural relations—i.e. the geometrical relations—of representatum (the mental model) and representandum (the world) there exists a strong correspondence. If we take Johnson-Laird literally, this can only mean that the mental model has to have a Euclidean structure, because it is a model in "three dimensions" of the physical world.[14] On the other hand, what we perceive visually is mediated through the model and our perception is Euclidean only in its entirety.

The inputs of the different depth cues via perception are apparently non-geometrical; consequently they do not lead in a straightforward way to a Euclidean structure. Therefore, a mechanism has to be assumed that allows for the construction of (nearly) Euclidean mental models on the basis of non-geometrical depth cues. As shown above, the modified weak fusion theory of Landy describes such a mechanism, which would explain the Euclidean nature of mental models. But whether it is reasonable to assume that there are cognitive processes similar to Landy's processes of modified weak fusion or not is an empirical question and thus should be tested empirically.

However, mental models in visual perception appear to be different from mental models used in text understanding or the carrying-out of logical inferences. The latter models are 'opaque'—i.e. they are 'visible before the mind's eye' and can be altered—while mental models in visual perception seem to be 'transparent' (cf. Held 2006). If a model is built up while

[14] For every three-dimensional layout of objects holds that it has either a coherent geometrical structure or not. In the second case it would not be reasonable to talk of a "strong correspondences" to the (locally) Euclidean-structured physical space at all. But if the model has a geometry, it has to be either Euclidean or Non-Euclidean. Voting for the second leaves the problem why perceptual space *in toto* is experienced as Euclidean still unsolved.

reading a text, this model might change any time. If, for example, some expectations resulting from the model are in conflict with new information provided by the text, the model will be changed in a way that the new information can be incorporated and so will be consistent again. There is always a feedback between the current model, expectations based on that model, and new information. Furthermore, all this can be made conscious. One can be aware of the model and of the changes it undergoes. One even can effect the changes deliberately.

In visual perception on the other hand it is not possible to compare the actual model with the incoming visual information and to 'play' with it in order to compare different alternatives, to judge their validity. Only in very exceptional cases—e.g. looking at a Necker Cube or the famous Rubin Vase—it seems to be impossible to reach a stable interpretation of the visual data. On the other hand not even the knowledge that one is presented an illusionary picture—e.g. the Müller-Lyer-Illusion—prevents one from being trapped in this illusion. Previous knowledge here is not sufficient to lead to the correct interpretation of the data.

These considerations lead to certain constraints on mental models used in perception. Contrary to those used in reasoning or text-understanding, mental models in perception cannot be modified consciously, nor are they the object of consciousness (i.e. they are opaque). Some (but not all) of the processes involved in the construction seem to be cognitively impenetrable. This would explain the fact, that visual illusions are stable even in the light of additional information. Furthermore, whereas mental models in text-understanding are representations of the content of the text, mental models in visual perception act like a filter through which we perceive the physical world.

The conclusion therefore takes the form of a conditional: If—as suggested by Johnson-Laird—mental models play a role in visual perception, then processes of model construction have to be assumed, which implement some kind of modified weak fusion. In addition, some of the processes are cognitively impenetrable. Presently, I have to leave open the question whether, in the light of these findings, it is still reasonable to talk of mental models—in the sense in which Johnson-Laird uses the term. However, new empirical data may lead to a refinement of the idea of mental models, and show how the models used in text-understanding and the ones used in visual perception are related.

5. Summary

Physical space as well as picture space has a Euclidean structure. Nine depth cues can be exploited while perceiving physical space, five of them while perceiving pictorial space. Information about the spatial layout of objects can be derived from the perception of a picture and from perceiving the real scene in a very similar way.

However, the question whether vision space has a geometrical structure at all seems to be open. If it has, it is still undecided whether it is Euclidean, hyperbolic or spherical. The findings of Zimmer suggest that there is no coherent geometry at all. But it seems certain that the more cues are available and the richer the environment gets the more vision space becomes Euclidean. It seems also clear, on the other hand, that single cues cannot be modelled by one of the standard geometries.

If mental models are assumed to be an essential part of visual perception, the question is how they are built up having a geometrical structure similar to the physical world, although the information used is non-geometrical. The specific process of model construction required may, in the light of the previous considerations, be best explained as a form of modified weak fusion. Furthermore, we have to assume that in visual perception some of the processes used in model construction are unconscious and cognitively impenetrable. They use visual data, which have no geometrical structure and present only ordinal information, and transform them into a full-fledged three-dimensional model of the world. Whether or not this conception of the perceptual process is still compatible with the theory of mental models, as presented by Johnson-Laird, remains to be shown.

References

Angell, R. B. (1974), 'The geometry of visibles', *Noûs* **8**, 87–117.

Blank, A. (1957), 'The geometry of vision', *British Journal of Physiological Optics* **14**, 154–169.

Blank, A. (1961), 'Curvature of binocular visual space: An experiment', *Journal of the Optical Society of America* **51**, 335–339.

Blumenfeld, W. (1913), 'Untersuchungen über die scheinbare Größe im Sehraume', *Zeitschrift für Psychologie* **65**, 241–404.

Burton, H. (1945), 'The optics of Euclid', *Journal of the Optical Society of America* **35**, 357–372.

Cutting, J. (2003), Reconceiving perceptual space, *in* H. Hecht, R. Schwarz & M. Atherton, eds, 'Looking Into Pictures', MIT-Press, Cambridge, MA.

Cutting, J. & Vishton, P. M. (1995), Perceiving layout and knowing distancces: The integration, relative potency, and contextual use of different information

about depth, *in* W. Epstein & S. Rogers, eds, 'Perception of Space and Motion', Academic Press, San Diego.

Daniels, N. (1972), 'Thomas Reid's discovery of a non-Euclidean geometry', *Philosophy of Science* **39**, 219–234.

Foley, J. (1972), 'The size-distance relation and intrinsic geometry of visual space', *Vision Research* **12**, 323–332.

Gibson, J. J. (1950), *The Perception of the Visual World*, Houghton Mifflin, Boston.

Gogel, W. (1963), 'The visual perception of size and distance', *Vision Research* **3**, 101–120.

Goldstein, E. B. (1996), *Sensation and Perception*, 4th edn, Brooks/Cole Publisher, Pacific Grove, CA.

Grünbaum, A. (1963), *Philosophical Problems of Space and Time*, Knopf, New York.

Hardy, L. H., Rand, G., Rittler, M. C., Blank, A. A. & Boeder, P. (1953), *The Geometry of Binocular Space Perception*, Schiller, Elizabeth, NJ.

Hecht, H., Van Doorn, A. & Koenderink, J. (1999), 'Compression of visual space in natural scenes and in their photographic counterparts', *Perception & Psychophysics* **61**, 1269–1286.

Held, C. (2006), Mental models as objectual representations, this volume, pp. 237–253.

Hillebrand, F. (1902), 'Theorie der scheinbaren Größe beim binokularen Sehen', *Denkschrift der Kaiserlichen Akademie der Wissenschaften Wien, Mathematisch-Naturwissenschaftliche Classe* **72**, 255–307.

Indow, T. (1967), 'Two interpretations of binocular visual space: Hyperbolic and Euclidean', *Annals of the Japanese Association for Philosophy of Science* **3**, 51–64.

Indow, T. (1974), 'On geometry of frameless binocular perceptual space', *Psychologia* **17**, 50–63.

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

Johnson-Laird, P. N. (1988), *The Computer and the Mind: An Introduction to Cognitive Science*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P. N. (1992), Mental models, *in* S. Shapiro, ed., 'Encyclopedia of Artificial Intelligence', 2nd edn, John Wiley, New York.

Johnson-Laird, P. N. (1998), Imagery, visualization, and thinking, *in* J. Hochberg, ed., 'Perception and Cognition at Century's End', Academic Press, San Diego.

Kant, I. (1998), Kritik der Reinen Vernunft, *in* W. Weischedel, ed., 'Werke in sechs Bänden', Wiss. Buchges., Darmstadt.

Kemp, M. (2001), *Leonardo on Painting*, Yale University Press, New Haven, London.

Landy, M. S. & Maloney, L. T. (1998), Combination of cues and priors in depth perception, *in* S. Saida & K. Sagawa, eds, 'Proceedings of the International Workshop on Visual Cognition', National Institute of Bioscience and Human-Technology, Tsukuba (Japan).

Landy, M. S., Maloney, L. T., Johnston, E. B. & Young, M. J. (1995), 'Mea-

surement and modeling of depth cue combination: In defense of weak fusion', *Vision Research* **35**, 389–412.

Luneburg, R. K. (1947), *Mathematical Analysis of Binocular Vision*, Princeton University Press, Princeton, NJ.

Luneburg, R. K. (1950), 'The metric of binocular space', *Journal of the Optical Society of America* **40**, 62–642.

Rehkämper, K. (2003a), *Bilder, Ähnlichkeit und Perspektive. Auf dem Weg zu einer neuen Theorie der bildhaften Repräsentation*, DUV, Wiesbaden.

Rehkämper, K. (2003b), What you see is what you get: The problems of linear perspective, *in* H. Hecht, R. Schwarz & M. Atherton, eds, 'Looking Into Pictures', MIT-Press, Cambridge, MA.

Reid, T. (2000), *An Inquiry into the Human Mind on the Principles of Common Sense*, Pennsylvania University Press, University Park.

Rogers, S. (1995), Perceiving pictorial space, *in* W. Epstein & S. Rogers, eds, 'Perception of Space and Motion', Academic Press, San Diego.

Rogers, S. (2003), Truth and meaning in pictorial space, *in* H. Hecht, R. Schwarz & M. Atherton, eds, 'Looking Into Pictures', MIT-Press, Cambridge, MA.

Schelling, H. (1956), 'Concept of distance in affine geometry and its applications in theories of vision', *Journal of the Optical Society of America* **46**, 309–315.

Strawson, P. (1996), *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*, Methuen, London.

Suppes, P. (1977), 'Is visual space Euclidean?', *Synthese* **35**, 397–421.

Suppes, P., Krantz, D., Luce, R. D. & Tversky, A. (1989), *Foundations of Measurement*, Vol. 2, Academic Press, New York.

Trudeau, R. (1985), *The Non-Euclidian Revolution*, 2nd edn, Birkhäuser Publishers, Boston.

Wagner, M. (1985), 'The metric of visual space: Perception & psychophysics', *Synthese* **38**, 483–495.

Wheatstone, C. (1838), 'On some remarkable, and hitherto unobserved, phenomena of binocular vision', *Phil. Trans. Roy. Soc.* **8**, 371–394.

Zimmer, K. (1998*a*), *Experimentelle Untersuchungen zur geometrischen Struktur des binokularen Sehraums*, Shaker Verlag, Aachen.

Zimmer, K. (1998*b*), Intrinsic geometry of binocular space, *in* Y. Lacouture & S. Gondin, eds, 'Fechner Day 98: Proceedings of the Fourteenth Annual Meeting of the International Society for Psychophysics', Quebec.

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Emotion, Decision, and Mental Models

Michael Pauen

Otto-von-Guericke-Universität Magdeburg [1]

Abstract
Recent research has shown that rational decisions may require the participation of emotions. It would follow that an adequate model of real-world decision-making has to account for emotions in some way or other. Due to their multi-modal character and because they preserve the structure of the objects or states of affairs they represent, mental models are particularly well-suited for this undertaking. Starting with a sketch of the underlying understanding of mental representation and emotion, the impact of emotion on cognition and decision will be outlined. These considerations provide the basis for a number of constraints for a theory of mental representation. It will turn out that, unlike a symbolic theory of mental representation, the theory of mental models does justice to these constraints.

The theory of mental models has proven fruitful mainly in the areas of deductive reasoning (Markovits & Barrouillet 2002, 410) arguing and rational decision-making (Johnson-Laird & Shafir 1993, Green 1996), and self-consciousness (Metzinger 1993). Although the multimodal character of mental models gives room for emotions as part of these representations, a systematic analysis of the role of emotion has yet to be developed, even if some aspects have already been discussed (Oatley & Johnson-Laird 1987, McCloy & Byrne 1999). Such an analysis is particularly desirable because evidence from psychology (Bless & Forgas 2000) and cognitive neuroscience (Lane & Nadel 2000, Damasio 1994, 1999, Bechara et al. 1997) demonstrates the crucial role of emotions in decision making and voluntary action.

[1] E-mail: m@pauen.com

While philosophers and psychologists alike have always conceded that emotions, affects, and moods actually do have an impact on decision-making , they used to think that emotions tend to impede or even obstruct reasonable decision-making. Recent research, by contrast, has shown that rational decisions may even *require* the participation of emotions (Bechara et al. 1997, Damasio 1994). Conversely, rational decisions may be impeded not by the *presence* but, rather, by the *absence* of emotions. It would follow that an adequate model of real-world decision-making has to account for emotions in some way or other.

I will argue that mental models are particularly well-suited for this undertaking; this is due to their multi-modal character and because they preserve the structure of the objects or states of affairs they represent. In doing so, I assume that something like the PDP-Model of the neural architecture and function (Rumelhart et al. 1986) is on the right track, that is, I assume that cognitive processes are based on the activity of parallel distributed processing of neural assemblies.

In the first section of the following paper, I will give a sketch of the underlying understanding of mental representation and emotion. The second part will then focus on the connection between emotion, cognition, and decision. These considerations provide the basis for a number of constraints for a theory of mental representation that I will derive in the third part. I will conclude that, unlike a symbolic theory of mental representation, the theory of mental models does justice to these constraints.

1. Mental representation, mental models, and emotion

1.1. Mental representation

Talking about mental models is talking about mental representation. I take representation to be an asymmetrical relation between a representation bearer and an (abstract or concrete) object that is represented. The relation is based on a—more or less complex—rule that serves as the basis for the interpretation of the representation bearers. A city-map may represent the streets of a city and part of the rule may be the scale or a color-code that indicates certain properties of streets, buildings, etc. Mental representations are specific not only because their representation bearers are mental states but also because they need no separate interpreter: The representation is interpreted by the very subject whose representation it is. While most philosophers would agree that mental representations have an intentional content, it is controversial whether this content can be "naturalized" such that the content of a representation can be determined on the basis

of non-semantic information, say about the causal history of the mental state in question (Pauen 1996, Stich & Laurence 1994). Another issue is whether the relation between a representation bearer and its content is merely symbolic and, therefore, arbitrary, at least in principle, or whether there is an "intrinsic" (Palmer 1978, 271) relation between content and representation such that similarities of certain object properties show up as similarities of the corresponding properties of the representation bearers. Many philosophers describe such a relation as "analogous" or structure preserving (Blachowicz 1997). Mental models belong to the latter category.

Finally, a theory of mental representation should say something about implementation on the physical level. The most interesting question is how semantics is connected to causality, that is how the content of mental representations causally affects the corresponding neural activity in the brain. Based on the idea that the mind works more or less like a traditional Von-Neumann-Computer, Jerry Fodor has developed a detailed answer to this problem from the perspective of a symbolical theory of mental representation. According to Fodor, mental representations provide the connection because their syntactical properties which determine their functional role in neural processing correspond to their semantic content:

> Mental representations can mediate the world's effects upon behavior because the same properties of mental representations that determine their computational roles also carry information about the world. More particularly computation is by definition syntactic, and information is by definition etiological, and mental representations can mediate between behavior and the world because their syntactic structure carries information about their ... causal histories. (Fodor 1994*a*, 86; compare Fodor 1994*b*)

One of the problems with this theory is that there is too much evidence that the mind doesn't work that way, that is, it does *not* work like a traditional computer (Edelman 1992, 152, 225; Rumelhart et al. 1986). It is even more important in the present context that, as I will demonstrate below, Fodor's theory cannot account for empirical evidence concerning the relation between emotion and cognition. The theory of mental models, by contrast, *does* account for these findings, but it should say something about the relation between semantics and causation.

1.2. Emotion

The idea that mental models might prove helpful for our understanding of the interaction between emotion and cognition is not particularly new. It has already been elaborated to some extent by Oatley & Johnson-Laird (1987). According to the authors, "emotions arise as disturbances which accompany interruptions and discrepancies among multiple goals and rep-

resentations" (Oatley & Johnson-Laird 1987, 30). Emotions are thought to be "part of a management system to co-ordinate each individual's multiple plans and goals under constraints of time and other limited resources" (Oatley & Johnson-Laird 1987, 31). The underlying assumption is that our cognitive system is made up of different modules whose operations need to be coordinated. Coordination may be propositional and symbolic, but in addition, Oatley and Johnson-Laird postulate the existence of a more primitive way of coordination and communication between the modules: "The other kind of communication is non-propositional. It is simpler, cruder, and evolutionarily older. Non-propositional signals have no internal symbolic structure of significance to the system. They do not denote anything. Like hormones, they function purely causally." (Oatley & Johnson-Laird 1987, 32) This includes that emotions only *prepare* the organism for an action. In order to develop fully, emotional experience has to include "a conscious evaluation of the juncture in planning, based on propositional signals reaching the operating system so that it is able to ascribe a meaning to the emotion mode, and so that voluntary action can be scheduled" (Oatley & Johnson-Laird 1987, 34). Following this model, emotions arise particularly at significant junctures in the execution of our plans, and they help to "organise a transition to a new phase of planned activity directed to the priorities of the mode with associated goals and certain stored plans for dealing with what has happened" (Oatley & Johnson-Laird 1987, 35).

I do not doubt that emotions play the roles that Oatley and Johnson-Laird mention, and I also agree that emotions are simpler and cruder than fully developed cognitive processes. What I doubt, however, is that this is an exhaustive description of the interaction between emotion and cognition. I do not believe, in particular, that emotions, taken by themselves, play *only* a simple causal role like hormones, such that their specific content, whatever it may be, stems from cognitive processes only.

Emotion has been an important issue throughout the history of philosophy,[2] but many philosophers used to think that emotions interfere with and impede truly rational decision and cognition. According to the Stoics, "the passions are diseases of the soul analogous to those of the body, and like the latter may be distinguished as to constitutional morbid propensity" (Gardiner et al. 1970, 66). Kant even compares affects to cancerous diseases and passions to tuberculosis.[3]

[2] See Gardiner et al. (1970).

[3] Kant (1902, Vol. VII, 266, 252). On the other hand, there is also the view that emotions are rational in a certain respect. Compare Crousaz (1715, 64), who thinks that emotional reactions are shortcuts to explicit rational assessments: "Tout ce donc qui faisant impression sur les organes de nos sens, quand ils ne sont point dérangez, donne lieu à des sentimens agréables, est fait & agit d'une manière dont l'idée nous plairoit déjà par elle-même, si nous en avions la connoissance."

Current research, by contrast, has provided a growing body of evidence that emotions do not only *support* rational, cognitive processes but that they are almost *indispensable* for rational decision and action. In general, emotions seem to play a "metacognitive" role (Bless & Forgas 2000), that is, they provide shortcuts to results that would be much more difficult to come by on the basis of cognitive operations only.

In the following sketch, I will try to evade the conflicts between the different theories of emotion, as far as this is possible. I assume that emotions are evolutionary determined patterns of psycho-somatic activity that motivate and prepare conducive cognition and behavior, particularly in situations that are critical for the survival and the reproduction of an organism (Damasio 2000).

From the first person perspective, emotions differ with respect to their strength or intensity and with respect to their quality (fear *vs.* joy). Typically, the quality can be said to be either aversive, indicating a "punisher" (fear, disgust), or positive, indicating a "reward" (joy), thus motivating or preparing a certain kind of behavior (Rolls 1998).

Like perceptions, most emotions have intentional content, that is, they are directed at or, simply, are "about" something: You may be afraid *of* an upcoming thunderstorm or you may be enthusiastic *about* meeting an old friend again. On the other hand, emotions differ from perceptions insofar as they tend to include a positive or negative *evaluation* of their referent. As long as you simply watch a certain picture, you may remain completely neutral. But if you react emotionally, then the distance vanishes and you take a stand: You may hate or love it, you may be afraid or become aggressive, you may want to get rid of it or possess it.[4]

Unlike cognitive states, emotions are not under the individual's conscious control. One cannot willingly stop feeling afraid. Finally, emotions differ from instincts because they are subject to learning and, therefore, highly flexible. If you have been attacked by a dog in a certain situation, chances are that you will feel afraid if you meet a dog in a similar situation again.

Fear is an instructive example. Fear is a negative emotion that indicates the presence of a dangerous object or a dangerous situation. According to the *Oxford English Dictionary*, fear is "the emotion of pain or uneasiness caused by the sense of impending danger, or by the prospect of some possible evil" (The Oxford English Dictionary 1989).

On closer inspection, several levels of description can be distinguished. From the first person perspective, fear states have a distinct qualitative character. It is almost impossible to give an adequate verbal description of this character but it is easily recognized from the first person perspective. Seen from a cognitive point of view, fear states imply an increase in atten-

[4] For a recent defense of this view compare Goldie (2000).

tion and a focus on the prospects of the current situation (Roseman et al. 1994). On the behavioral level, fear states are associated with avoidance and protective behavior, that is, with the readiness for an action that can reduce the possibility of harm (Frijda et al. 1989, Öhman 1993, Roseman et al. 1994, Meyer et al. 1997). Finally, fear states involve certain somatic reactions including a particular type of arousal, cardiac responses, deep breathing, and certain reactions from the endocrine system (Öhman 1993, Roseman et al. 1994, LeDoux 1996).

2. Emotion, decision, and cognition

Particularly, cognitive theories of emotion assert that emotions have a strong cognitive component; an "appraisal" that is of fundamental relevance for the way an emotion is experienced (Frijda et al. 1989, Frijda 1993). As I have already outlined above, recent research has provided converging evidence that the connection between emotion and cognition also works the other way round: Emotion has an impact on cognitive processes. In what follows, I will present some empirical results that support this view.

2.1. Somatic markers

The interaction between emotion and cognition takes center stage in the work of Damasio whose "Somatic Marker Hypothesis" has been among the most influential theories of emotion in recent years. According to Damasio (1994, 1999), somatic markers are emotional reactions with a strong somatic component that support decision making, including rational decision making. These reactions are based upon the individual's previous experiences with similar situations. Somatic markers permit a comparatively fast preselection of the relevant alternatives which are then subjected to a more detailed cognitive processing for the final decision. In doing so, somatic makers increase the efficiency and accuracy of human decision making. Following Damasio, decision making would be almost impossible if detailed cognitive processing of *all* the available alternatives were necessary.

Damasio refers to several case studies and experiments that seem to show that the inability to experience emotions results in a severe impairment of rational decision-making. In an experiment conducted by Bechara et al. (1997), healthy controls and patients with emotional deficits had to perform a gambling task which required a rational decision for the most advantageous strategy in order to gain as much money as possible. The controls started with an emotional reaction, then they adopted the advantageous strategy before they were finally able to tell what the advantageous strategy

was, a few trials later. Patients, by contrast, showed no emotional reaction and continued to use the disadvantageous strategy throughout the experiment, although they also realized what the advantageous strategy was. The experiments support the basic idea underlying the somatic marker hypothesis, namely that rational decision making requires emotional reactions.

Note that these experiments seem to indicate also that, in contrast to the hypothesis of Oatley & Johnson-Laird (1987), emotional reaction is an ongoing process that does not require unexpected events or specific junctions in the proceeding of our plans. Second, in order to perform the function described, emotions have to be very specific, and third, they have to acquire this specificity independently from cognitive processes: In the above experiment, the cognitive assessment follows only a while *after* the emotional reaction has set in.

2.2. Affect infusion

Similar conclusions can be drawn from the work of Forgas (1995) and Forgas et al. (2000). The authors postulate a single, automatic, and basically unconscious "Mood-Management System" that controls the interaction between cognitive and emotional processes. This means, first, that the system is involved in the cognitive control of emotions. It monitors emotions and moods continuously in the background and becomes active if emotions, particularly negative ones, pass a certain threshold. Second, affects may exert a direct influence on decision and cognition, in particular when open constructive processing is required. "Affect as information" is a heuristic strategy. Individuals ask themselves "How do I feel about it?," especially if processing resources are limited, the task is not relevant, and there is no prior knowledge. "Affect-priming," by contrast, is more likely in the case of substantive processing. In this case, a certain emotion gives access to additional information in the individual's memory that is associated with this emotion.

2.3. Feeling of knowing, availability heuristics, and the "sample size effect"

In addition to the *direct* impact on cognition and decision outlined so far, emotions seem to have also an *indirect* impact on our choice of cognitive strategies which, in turn, lead to certain decisions. In general, positive moods seem to correlate with non-analytical, "top-down" problem-solving strategies while negative moods correlate with analytical "bottom-up" processing. But why is this so? Garcia-Marques & Mackie (2000) refer to another well known effect: People tend to rely on non-analytical top-down

processing also when they feel familiar with a problem such that they have a "feeling of knowing." The explanation, then, is that individuals attribute their positive mood to their familiarity with a problem and choose the non-analytical strategy because this strategy would be adequate for a familiar problem.

The idea that emotions guide our choice of cognitive strategies gets support from two other phenomena. According to the "*availability heuristics*" (Tversky & Kahnemann 1973), individuals tend to use the ease of retrieval of a certain reason as a cue for its relevance in support of a hypothesis. In general, a reason that can be easily retrieved will be regarded as more relevant than a reason whose retrieval is difficult. As a consequence, someone who imagines only a few reasons in support of a certain hypothesis may become more convinced of that hypothesis than a person who imagines a larger number of reasons, say 6 rather than 3. The explanation is that it is easier to retrieve 3 reasons than to retrieve 6 (Haddock 2000).

Finally, when individuals have to make a choice between two options in order to achieve a certain result, their choice is apparently guided by the feeling of familiarity concerning the likelihood of the desired outcome in the two scenarios. Again, the feeling seems to be based on the ease of imagining the situation. If it is easier to imagine drawing a winning ticket from a large bowl than to draw it from a small bowl then we will feel more familiar with this situation and decide to take a ticket from the large bowl—even if the ratio between winning and blank tickets is identical in both cases (Brendl 2000).

2.4. Contextual conditioning

Given that one's access to a certain situation may imply information from different modalities, it might be asked whether emotional associations are restricted to a particular modality. Imagine that you have *seen* a certain scenario several times and now you *hear* about the same type of scenario. The question is, whether we would have to expect that you will feel familiar with this scenario. Research from other areas of the psychology of emotion indicates that this is so, due to so-called "contextual conditioning." [5] If you ran into a fearful situation, say on a dark desert highway, chances are that fear will be associated with almost all the features, that you experienced in this situation (warm smell of colitas, cool wind in your hair, hearing a mission bell). As a consequence, the experience of fear can be evoked by cues from other modalities.

[5] See LeDoux (1996, 165-169).

3. Constraints

The upshot of these observations is as follows: First, there is an obvious interaction between emotion and cognition, second, it seems that this interaction is based on associations between the present situation that asks for a decision or solution and similar situations in the past, and third, associations may go across different modalities, that is, visual as well as acoustic perception may be able to elicit an emotional response that affects cognitive processing of a certain situation.

It would follow that the crucial questions from the viewpoint of a theory of mental representation are: (1) What are the features that such a theory has to postulate in order to explain these functions and (2) do mental models have the required features?

I assume that four features are of special importance.

3.1. Similar input—similar output

Similar representations have to be able to activate each other. If the feeling of familiarity is supposed to give a reliable cue concerning my experience with a certain type of situation, then representations of this type should be activated by similar representations and *only* by similar actual representations. Thus, if I have experienced the large-bowl situation previously, then current representations of large-bowl scenarios should reliably activate representations of my previous experiences—otherwise either I would not be able to make use of my experience, or other representations might be activated that may be useless or even misleading. In either case, emotional experience would fail to enhance cognitive processing.

But how can this constraint be met? In order to discuss this point, it is necessary to refer to the distinction between digital/symbolic and analog representation that was already mentioned at the beginning of this paper. It has been argued that the difference can be defined either with respect to (a) resemblance between representation and referent, (b) continuity in analog representation *vs.* discontinuity in digital representation, and finally, (c) density in analog vs. non-density in digital representation.

Unfortunately, all these suggestions are subject to serious objections (Blachowicz 1997). (a) This is true in particular for the alleged resemblance between analog representations and their referents, as far as mental representations are concerned. Of course, the relation between a mental representation and its referent differs in principle from the relation between, say, a photograph and its object. If this were not so, then the best representation of an odor in the external world would be the replication of the odor in the mind, which is obviously nonsense. Trivially, perception-

based representations require a *transformation* of patterns of external stimuli into a mental representation and the underlying neural activity. It will turn out that analog representation requires that the structure of the referent is preserved in some way, but this structure-preservation must not be confused with our commonsense understanding of resemblance. (b) Discontinuity or discreteness is a necessary, but not a sufficient condition for digital/symbolic representation. It is not a sufficient condition because, as Blachowicz remarks, television and newspaper pictures are not continuous although they are certainly analog representations. (c) The same is true for density. According to Goodman (1976, 160), "a system is analog if syntactically and semantically dense." Being syntactically dense means that a scheme "provides for infinitely many characters so ordered that between each two there is a third" (Goodman 1976, 136). Again, the definition is open to counterexamples. A number of bar graphs that represent the development of a company's profit over the years are certainly analog representations, although they are not syntactically dense in Goodman's terms.

The reason why bar graphs count as analog representations is that they preserve certain relations or structures of their referents. It has been argued that the ability to preserve structure is the most relevant feature of analog representation. Blachowicz (1997, 74) defines it as "relational identity" which may hold on a quantitative as well as on a qualitative level. Relational identity on the *quantitative* level means that the relation between certain values in the referent world is identical with the relation of the same values in the representing world. Relational identity on the *qualitative* level is much more difficult to define. The comparison between the captain of a ship and the president of a country might serve as an example for qualitative identity: In both cases the relation is leadership.

A suggestion quite similar to Blachowicz's has been made by Palmer (1978), who points to the distinction between first- and second-order isomorphisms. While first-order isomorphisms require that the representing properties themselves are *identical* with the properties they represent, second-order isomorphisms require only an identity of the *relations* that hold between the properties:

> A representation is second-order isomorphic to its referent world if the similarity of represented objects is functionally reflected by the similarity of the corresponding representing objects." (Palmer 1978, 292)

This difference is of fundamental importance to the problem of similar input/similar output. If analog representations are relationally identical to their referents then they preserve the structure of those referents: The bar-graphs in the above example preserve the relation between the profits over the years such that similar profits result in similar graphs and different profits in different graphs. Propositional or symbolic representation cannot

guarantee this because symbolic representation permits arbitrary relations between symbols and referents.

Structure preservation is one of the most fundamental contentions of mental model theory: "The structures of mental models are identical to the structures of the states of affairs, whether perceived or conceived, that the models represent." (Johnson-Laird 1983, 419)

Taken by itself, this claim may appear as mere hand waving, as long as it is not explained *how* exactly mental models achieve structure-preservation. One way to come by such an explanation would be to look at the implementation, that is, to look at how analog representations may be realized on the neural level. I have already said why we need an answer to this problem. One reason is that at least one of the competing theories, namely Fodor's, implies a fairly precise idea of how symbolic mental representation might be implemented (Fodor 1994*a*).

I think, however, that mental model theory can avail itself of neural network theories like the Parallel-Distributed-Processing model (Rumelhart et al. 1986) in order to explain the implementation of analog representation. According to the PDP-theory, representations are instantiated as distributed patterns of activity. Different representations do not require different units, say artificial or organic neurons, but are realized as different patterns of activity that may involve the same units (Hinton et al. 1986, Rumelhart & McClelland 1986). Conversely, similar representations involve similar patterns of activity. The model is able to explain the very fact that a current representation can activate similar previous representations, just because the pattern underlying the current representation is similar to the patterns of similar previous representations. That's why the current representation tends to activate almost exactly those nodes that underlie the previous representations. Another important consequence is that neural networks permit "spontaneous generalizations" based on the similarity of the items that are subject to the generalization. So if you learn that chimpanzees like onions, your expectation that similar animals, say gorillas or orang-outans, like onions will rise. It is, again, the idea that similar representations are based on similar activity patterns that enables the PDP-theory to account for this finding. Given that the activity-patterns are similar such that they involve the same nodes, a connection between *one* of these patterns and a representation of onions will strengthen also the connection between the other patterns and the representation in question (Hinton et al. 1986).

All this is obviously a simplified picture, but it does show how the idea of analog representation can be implemented. So if the current situation is similar to some of your previous experiences, the respective representations will be similar, too. If you have experienced the large-bowl situation previously, chances are that your representations of the current and the previous

situation will preserve the similarity that is needed in order to explain the emotional reactions under discussion. This is a clear advantage compared to symbolic representation, which does not preserve structure and therefore fails to meet the present constraint.

3.2. Dynamic representations

Typically, problem-solving situations are dynamic. They require an interaction of the individual with some external conditions such that these conditions are changed in a certain way. This is true already for simple problems like making a choice between a large and a small bowl, and it should be true *a fortiori* for more complex problem-solving or decision-making scenarios. If the individual wants to make use of previous experiences, it seems necessary that these experiences are represented such that the connection between a cause, probably an action, and an effect, is represented, and this requires dynamic representation. I take it as obvious that this is true particularly if we want to explain how our feeling concerning a situation can have an impact on our cognitive processing: Asking yourself "How do I feel about it" would not make very much sense if there were no reliable connection between a present and a past dynamic scenario.

This constraint may seem trivial because the dynamic character of mental representation is so evident from the viewpoint of first person experience. That's why we are able to perform mental simulations of possible actions in order to detect advantages or disadvantages of the alternatives at hand without being obliged to execute these actions in the real world.

According to Johnson-Laird, mental models meet this constraint, given that they "can represent ... the temporal or causal relations between events. ... Models have a content and form that fits them to their purpose, whether it be to explain, to predict, or to control" (Johnson-Laird 1983, 410). Again, the important point is that these models are structure-preserving with respect to their referents. In the present case, the relevant structure is the relation between cause and effect. Thus, if I have found out that certain objects react in a distinctive way, it is not a bad guess that a present object of this sort, that is, a similar object will show a similar effect, too. Conversely, it is difficult to see how symbolic theories of representation would handle cases like this. Of course, cause-effect relations might be stored explicitly, but this would have to be done for every situation type. This seems to be somewhat expensive, and I doubt that it can be done at all. But even if it can, there will be no place for emotion in this picture: Subsuming a certain situation token under the related type would give you direct access to all the information that belongs to this type; even

if there was an interaction between cognition and emotion there would be no additional information that it could bestow you with.

3.3. Multi-modality

If an emotional response is associated with a certain scenario, the association seems to comprise cues from different modalities that were part of the scenario, such that each of these cues can recall the memory. It is obvious that this sort of "contextual conditioning" requires multimodal representation of various aspects of the scenario in question. Imagine that you once were attacked in a dark forest, then your ability to react fearfully to a contextual stimulus, say a sound, requires that the sound has been stored as part of the fear-evoking scenario. I conclude that in order to account for the above findings concerning the interaction of emotion and cognition and in particular concerning the relevance of contextual stimuli for the evocation of emotional responses, an adequate theory of mental representation has to be multimodal, that is, it has to comprise the representation of stimuli from different modalities. While it is difficult to see how this constraint could be satisfied by a symbolic theory, mental models meet it because of their multimodal character. Mental models, that is, "play a central and unifying role in representing objects, states of affairs, sequences of events, the way the world is, and the social and psychological actions of daily life" (Johnson-Laird 1983, 397).

4. Conclusion

It would seem, then, that mental model theory meets all the relevant constraints that can be derived from the above findings concerning the interaction of emotion and cognition. This is a serious advantage compared to familiar symbolic theories of mental representation. The advantage is of particular importance, given that ongoing research is providing us with ever more insights into emotion in general and into the interaction between emotion and cognition in particular.

All this does, of course, not prove that mental model theory is true. It may well be that the theory fails to do justice to other important constraints. An even more serious objection is that the idea of mental models is highly metaphorical. While we have a fairly precise idea of, say architectural and scientific models, our understanding of mental models is based, first, on the analogy with those models and, second, on the distinction from propositional or symbolic forms of mental representation. Third and most importantly, constraints that can be derived from scientific findings

may help to improve our understanding of mental models. That's why I think that the metaphorical character of mental models is not a real disadvantage. The details can and have to be filled in as our knowledge about cognitive processes grows. What we need is a framework that helps us to make sense of the empirical details and I think that this is what mental model theory does.

References

Bechara, A., Damasio, H., Tranel, D. & Damasio, A. R. (1997), 'Deciding advantageously before knowing the advantageous strategy', *Science* **275**, 1293–1295.

Blachowicz, J. (1997), 'Analog representation beyond mental imagery', *The Journal of Philosophy* **94**, 55–84.

Bless, H. & Forgas, J. P. (2000), The message within: Toward a social psychology of subjective experience, *in* H. Bless & J. P. Forgas, eds, 'The Message Within: The Role of Subjective Experience in Social Cognition and Behavior', Psychology Press, Philadelphia.

Brendl, C. M. (2000), Subjective experience and the effect of sample size on likelihood judgments, *in* H. Bless & J. P. Forgas, eds, 'The Message Within: The Role of Subjective Experience in Social Cognition and Behavior', Psychology Press, Philadelphia.

Crousaz, J. P. d. (1715), *Traité du beau*, Amsterdam.

Damasio, A. R. (1994), *Descartes' Error: Emotion, Reason, and the Human Brain*, Putnam, New York.

Damasio, A. R. (1999), *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, Harcourt Brace, New York, San Diego, London.

Damasio, A. R. (2000), A second chance for emotion, *in* R. D. Lane & L. Nadel, eds, 'Cognitive Neuroscience of Emotion', Oxford University Press, New York, Oxford.

Edelman, G. M. (1992), *Bright Air, Brilliant Fire: On the Matter of the Mind*, Basic Books, New York.

Fodor, J. A. (1994*a*), *The Elm and the Expert: Mentalese and Its Semantics*, MIT Press, Cambridge, MA, London.

Fodor, J. A. (1994*b*), Fodor's guide to mental representation: The intelligent auntie's vade-mecum, *in* S. P. Stich & T. A. Warfield, eds, 'Mental Representation: A Reader', Blackwell, Oxford.

Forgas, J. P. (1995), 'Mood and judgment: The affect infusion model (aim)', *Psychological Bulletin* **117**, 39–66.

Forgas, J. P., Ciarrochi, J. & Moylan, S. (2000), Subjective experience and mood regulation: The role of information processing strategies, *in* H. Bless & J. P. Forgas, eds, 'The Message Within: The Role of Subjective Experience in Social Cognition and Behavior', Psychology Press, Philadelphia.

Frijda, N. H. (1993), Moods, emotion episodes, and emotions, *in* M. Lewis & J. M. Haviland, eds, 'Handbook of Emotions', The Guilford Press, New York,

London.

Frijda, N. H., Kuipers, P. & ter Schure, E. (1989), 'Relations among emotion, appraisal, and emotional action readiness', *Journal of Personality and Social Psychology* **57**, 212–228.

Garcia-Marques, T. & Mackie, D. M. (2000), The positive feeling of familiarity: Mood as an information processing regulation mechanism, *in* H. Bless & J. P. Forgas, eds, 'The Message Within: The Role of Subjective Experience in Social Cognition and Behavior', Psychology Press, Philadelphia.

Gardiner, H. N., Metcalf, R. C. & Beebe-Center, J. G., eds (1970), *Feeling and Emotion, a History of Theories*, Greenword Press, Westport.

Goldie, P. (2000), *The Emotions: A Philosophical Exploration*, Clarendon Press, Oxford.

Goodman, N. (1976), *Languages of Art: An Approach to a Theory of Symbols*, 2nd edn, Hackett, Indianapolis.

Green, D. (1996), Models, arguments, and decisions, *in* J. Oakhill & A. Garnham, eds, 'Mental Models in Cognitive Science: A Festschrift for Philip Johnson-Laird', LEA, London.

Haddock, G. (2000), Subjective ease of retrieval and attitude-relevant judgments, *in* H. Bless & J. P. Forgas, eds, 'The Message Within: The Role of Subjective Experience in Social Cognition and Behavior', Psychology Press, Philadelphia.

Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. (1986), Distributed representations, *in* D. E. Rumelhart & J. L. McClelland, eds, 'Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations', MIT Press, Cambridge, MA.

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

Johnson-Laird, P. N. & Shafir, E. (1993), 'The interaction between reasoning and decision making: An introduction', *Cognition* **19**, 1–9.

Kant, I. (1902), *Gesammelte Schriften*, Berlin.

Lane, R. D. & Nadel, L. (2000), *Cognitive Neuroscience of Emotion*, Oxford University Press, Oxford.

LeDoux, J. (1996), *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*, Touchstone, New York.

Markovits, H. & Barrouillet, P. (2002), 'The development of conditional reasoning: A mental model account', *Developmental Review* **22**, 5–36.

McCloy, R. & Byrne, R. M. (1999), Thinking about what might have been: If only, even if, causality and emotions, *in* M. Hahn & S. C. Stoness, eds, 'Proceedings of the Twenty First Annual Conference of the Cognitive Science Society', Lawrence Erlbaum, London.

Metzinger, T. (1993), *Subjekt und Selbstmodell. Die Perspektivität phänomenalen Bewußtseins vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentation*, Schöningh, Paderborn, München, Wien.

Meyer, W.-U., Schützwohl, A. & Reisenzein, R. (1997), *Einführung in die Emotionspsychologie. Evolutionspsychologische Emotionstheorien*, Vol. 2, Huber, Bern, Göttingen.

Oatley, K. & Johnson-Laird, P. (1987), 'Towards a cognitive science of emotion', *Cognition and Emotion* **1**, 29–50.

Öhman, A. (1993), Fear and anxiety as emotional phenomena: Clinical phenomenology, evolutionary perspectives, and information-processing mechanisms, *in* M. Lewis & J. M. Haviland, eds, 'Handbook of Emotions', The Guilford Press, New York, London.

Palmer, S. E. (1978), Fundamental aspects of cognitive representation, *in* E. Rosch & B. B. Lloyd, eds, 'Cognition and Categorization', L. Erlbaum Associates, Hillsdale, NJ.

Pauen, M. (1996), 'Wahrnehmung und mentale Repräsentation', *Philosophische Rundschau* **43**, 243–64.

Rolls, E. T. (1998), *The Brain and Emotion*, Oxford University Press, Oxford.

Roseman, I. J., Wiest, C. & Swartz, T. S. (1994), 'Phenomenology, behaviors, and goals differentiate discrete emotions', *Journal of Personality and Social Psychology* **67**, 206–221.

Rumelhart, D. E., Hinton, G. E. & McClelland, J. L. (1986), A general framework for parallel distributed processing, *in* D. E. Rumelhart & J. L. McClelland, eds, 'Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations', MIT Press, Cambridge, MA.

Rumelhart, D. E. & McClelland, J. L., eds (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, MIT, Cambridge, MA, London.

Stich, S. P. & Laurence, S. (1994), 'Intentionality and naturalism', *Midwest Studies* **19**, 159–182.

The Oxford English Dictionary (1989), 2nd edn, Oxford University Press, Oxford, New York.

Tversky, A. & Kahnemann, D. (1973), 'Availability: A heuristic for judging frequency and probability', *Cognitive Psychology* **5**, 207–232.

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Language Processing: Construction of Mental Models or More?

Barbara Hemforth [a] and Lars Konieczny [b]

[a]Laboratoire Parole et Langage, Université de Provence [1]
[b]Center for Cognitive Science, Universität of Freiburg

Abstract

Language comprehension is often seen as the incremental update of a mental model of the situation described in the text. With every word a reader or listener processes, the model is adjusted to fit the linguistic input. This conception of language comprehension sounds very plausible at first sight but its scientific utility strongly depends on the definition of a mental model. In this paper, we will discuss different definitions of mental models for text comprehension, as well as the way these models may have an impact on comprehension processes.

A second distinction we will discuss is an "eliminative" view of mental modeling compared to a "hybrid" view. An eliminative view denies the relevance of more linguistic levels of processing below the construction of a mental model; claiming that most if not all processing phenomena can be explained without reference to syntactic or semantic levels of representation. From a more hybrid perspective, inspection of a mental model is one among several factors influencing linguistic processing. We will argue for this latter perspective. (We are, of course, not arguing against every possible variety of an eliminative view, but just against some that have been and still are very prominent in the psycholinguistic literature. So there surely are conceivable variants that are compatible with our objections.)

[1] E-mail: barbara.hemforth@lpl.univ-aix.fr

1. Definition of mental models

Mental models are a constructs used not only for language related processes but in cognitive psychology in general. One of the most central research areas is the construction of mental models for spatial reasoning (see e.g. Vosgerau 2006, Vandierendonck et al. 2006). But even in its most central area of application, it is not fully clear what a mental model really is supposed to be. For example, in the original framework of mental model theory (Johnson-Laird 1983), mental models are a medium for mental representation. They are a tools to manipulate mental objects in order to arrive at solutions to problems. This interpretation mainly refers to the *modelling as processing* aspect of mental models. On the other hand, mental models are themselves handled as abstract mental objects that can be manipulated. From this perspective, they represent *the result of the modelling process.* It is mostly this latter view that is adopted in theories of human language processing.

When referred to in language processing, mental models are very often defined as mental constructs describing the knowledge a person has about a particular domain of the world (see e.g. Gernsbacher 1991). This definition is more or less equivalent to a general concept of background knowledge and rarely more specified than that. Experimental evidence showing that "mental models" exert an influence on language processing amounts more or less to evidence on the general relevance of world knowledge. This is surely the most general, least debatable, and thus least helpful definition. What a theory of mental models should really give us, is a tool to represent or even formalize the way background knowledge is applied in human language processing.

A more text-oriented version of this definition is proposed by Garnham (1985), who considers text comprehension as a process of constructing a model of the situation the text is about (be it real or imaginary). The construction of this model serves the linking of information in different parts of the text. This definition comes closer to the description of a discourse model as it is described in more linguistic theories (e.g. Kamp 1981, Kamp & Reyle 1993, Heim 1983). The situation models are, however, enriched with background knowledge that is not explicitly mentioned in the text (see also Van Dijk & Kintsch 1983). A central point here is that the mental model contains objects and relations referred to in the text, but not the linguistic structure (words and sentences) of the text itself. The linguistic representation, however, is available as a separate representation, the so-called text basis.

2. An example of a mental model

What is the basic idea of constructing a mental model of a situation described in a text? We will present a classic example to give a more detailed idea of the processes and representations involved. While reading a text, readers attempt to incrementally build up a non-linguistic representation of the situation described. If the description is unambiguous, they succeed in doing so and the model is directly accessible, such that possible implications or inferences can be read off.

Assume the following description:

Imagine a table with the following objects:

A fork that is on the left side of a plate.
A knife that is on the right side of the plate.
A glass that is behind the knife.

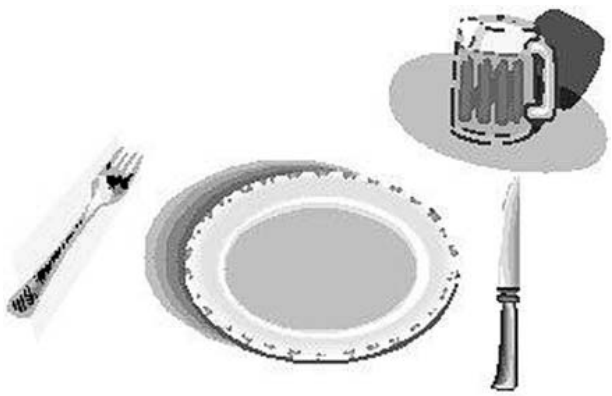


Fig. 1.

Given that you construct a mental model roughly equivalent to Figure 1, you can directly read off the answers to questions like: "Is the fork on the left side of the knife?" or "Is the glass on the right side of the fork?"

3. Arguments for direct effects of mental models in language comprehension

Since the early 1980s, many experiments have been conducted in the framework of mental models theory to show the impact of non-linguistic representations on language processing. We will describe a classic example by Glenberg et al. (1987) and a more recent one by Kelter et al. (2004) to give an idea of the logic behind these experiments.

In both examples, the basic idea is to keep the linguistic structure of the text as constant as possible. The only difference between conditions is due to the structure of the mental model. Let's first look at Glenberg's classic experiments on anaphor resolution. In their experiments, Glenberg and colleagues presented texts like (1). Participants were either presented with version (a) or with version (b) of the second sentence. When they were presented with the probe word "rose" after the fourth sentence, access was easier for the (a)-versions than for the (b)-versions of the materials. Note that the only difference between these conditions is the predicate that either implies that the rose is still with John (a) or that it is not (b). Between the two texts there is no difference with respect to linguistic complexity or the distance between the probe and its mentioning in the text. The most plausible explanation of this effect is that some kind of spatial distance between the current main protagonist (John) and the respective object (rose) affects accessability.

(1) 1. John was preparing for a date in the evening.
2. After dressing up,
(a) he grabbed **the rose** he had bought and went off.
(b) he forgot **the rose** he had bought and went off.
3. He took his sportscar to drive to his date.
4. He arrived a little late.
Probe: The **rose**...

A similarly intriguing example has recently been presented by Kelter et al. (2004). They showed that the temporal duration of events in a narrative affects accessability of referents. Participants in their experiments read texts like (2) (translated from German).

(2) *Setting:*
Mrs. Quasten is full heartedly concerned about her family's well-being. This year, as every year, she takes special care in arranging for New Years eve.
First event:

After getting up, she gets the **carp** ready for cooking and prepares the sauce.
Intermediate event:
Short: She then goes to the hairdresser and buys hairspray.
Long: She then goes to the hairdresser and gets a perm.
Third event:
When leaving the hairdresser, she hails a cab.
Probe: **carp**

When participants had to decide whether the probe word (carp) had been mentioned in the text, they took less time when the "real" time of the intermediate event was short than when it was long. Again, it is virtually impossible to find a purely text-based explanation of this effect. The temporal structure of the situation described in the text affected the accessibility of the referent for the probe.

Given the experiments presented here, there is no doubt that non-linguistic representations of a text play a role in text comprehension. We only gave a very small sample of the evidence available (for a more general overview, see Garrod forthcoming). However, the evidence is compatible with two perspectives on the role of mental models in language processing: a hybrid view where the situation model is one of the representations affecting processing, leaving room for more linguistic levels of representation as well, and an eliminative view, where all processing phenomena are finally based on non-linguistic representations, possibly even eliminating the need for such intermediate representations.

4. Eliminative mental model Theory: The role of parsimony

Since Bever (1970), principles governing human sentence processing have been investigated by looking at a very specific type of sentences, so-called garden-path sentences. These are sentences containing some kind of local ambiguity that is initially interpreted in a way incompatible with semantic or syntactic information showing up later in the sentence. A classic example of a garden-path sentence is (3).

(3) The horse raced past the barn fell.

Initially this sentence is rated as ungrammatical even by most native speakers. The problem is that there is a local ambiguity on the word "raced," which may be the main verb of the sentence, or a past participle starting

a reduced relative clause (... that was raced past the barn ...). There is a strong preference to interpret "raced" as the main verb of the matrix clause, but then the verb "fell" cannot be integrated, giving the impression of ungrammaticality.

Since Frazier & Fodor (1978), or even since Kimball (1973), garden-path effects have been explained on the basis of the syntactic structure of the respective sentences. In (3), the main verb reading is syntactically less complex than the reduced relative clause so that any principle involving a minimal amount of structure (e.g. minimal attachment, Frazier 1987, or simplicity, Gorrel 1995) can predict the preferences.

A principle of parsimony in syntactic structure building can also explain preferences in constructions like (4i, ii).

(4) (i) The psychologist told the woman that he had problems with
a. her husband
b. to leave her husband.
(ii) The cop watched the spy with the binoculars.

In (4i) the (a)-version is syntactically less complex and easier to understand than the (b)-version. The ambiguity lies in the interpretation of the "that"-clause either as a complement clause attached to the main verb (He told her that ...) or as a relative clause modifying the object noun phrase (... the woman that ...). (4ii) is fully ambiguous (the prepositional phrase "with the binoculars" can be interpreted as an instrument of watching or as an attribute of the spy) but there is a preference to interpret the phrase "with the binoculars" as an instrument of the verb. This attachment is often assumed to be less complex (e.g. Frazier 1987; but see Hemforth 1993 for a critical analysis of this assumption).

Crain & Steedman (1985) proposed that the preferences established for these ambiguities are not at all due to a preference for a simple syntactic structure but to a simple and parsimonious mental model. During parsing, syntactic analyses are pursued in parallel. However, in cases of ambiguity, only those are kept which are compatible with the most parsimonious mental model. What does parsimony mean for a mental model? According to Crain and Steedman there are at least two factors determining more or less complex models. One is the number of presuppositions that have to be adjusted when integrating a new piece of information. For example, a definite noun phrase like "the woman" in (4i) implies that there exists a uniquely identifiable woman in the discourse universe. "The woman" in a context where there is more than one woman is infelicitous. A second factor concerning complexity is the number of entities to be represented in a mental model. Increasing the number of entities increases the complexity of the model. A preference for parsimonious models guarantees the con-

struction of mental models with the minimal number of entities necessary to represent the current discourse.

How can such an approach explain garden-path phenomena? Consider a discourse as in (5). In the target sentence the "that"-clause can be a complement clause ("told ... that") or a relative clause ("the woman that"). In the context given here, there is one married couple, presumably a man and a woman. Hence, the definite noun phrase "the woman" finds a unique referent without any further need of modification. Since there is no need for a modifying relative clause, the that clause is interpreted as a complement clause.

(5) Context:
A psychologist was counseling **a married couple**. One of them was nice to him, but the other one was fighting with him.
Target:
The psychologist told **the woman** that he was having trouble with
a. her husband.
b. to leave her husband.

What if there were two married couples and consequently two women in the context as in (6)? Then the definite noun phrase "the woman" would not on its own allow identification of a unique referent. There would be need for more information. In this case a modifying relative clause would be felicitous. In their experiments, Crain and Steedman found a preference for the complement reading in one-referent contexts like (5) and a preference for the relative clause reading in contexts like (6).

(6) Context:
A psychologist was counselling **two married couples**. One of them was nice to him, but the other one was fighting with him.

Out of any context, readers seem to prefer the complement clause reading (Frazier 1987). Why should that be so? Crain and Steedman assume that readers construct the minimal mental model compatible with the linguistic input. When they read "the woman," they assume that there is only a single woman in the current universe of discourse. Hence, there is no need for further information, no need for a modifying relative clause.

This line of research inspired an enormous amount of follow-up studies (e.g. Altmann & Steedman 1988), some of them pointing out empirical problems with the approach. It has been shown, for example that preferences for sentences like (4i) or (4ii) often depend on lexical biases (Britt et al. 1992), mostly from the verb, and also on syntactic biases like a preference for simple structures (Desmet et al. 2002, Konieczny & Voelker 2000).

In particular, lexical and syntactic effects show up as early processes, as they can be established in experiments registering eye movements while participants are reading texts. A more or less complex context, on the other side, only affects interpretation at later stages.

More importantly, Konieczny and Voelker showed that some of the predictions to be derived from Crain and Steedman's model don't really hold in all cases. They presented participants with short texts like (7).

(7) Yesterday, **a girl / two girls was / were** sitting on a bench.
Darja admired the girl a. with the pink dress. / b. with big eyes.

The interesting point here is that a fairly late preference for a noun modifying prepositional phrase ("the girl with the pink dress") could be established for the two-referent context.[2] However, there was no penalty for a noun phrase modifier in a one-referent context. Participants did not really care whether or not a definite noun phrase like "the girl" that was uniquely identifiable without further information was then modified by a prepositional phrase. This additional information is apparently easily acceptable even when it does not serve any referential purpose. But if it is not the case that readers have to update their mental model to a more complex one in cases of modified noun phrases, this approach cannot really explain the basic preferences established for isolated sentences.

A more recent version of the eliminative approach can be found in research applying the visual world paradigm (e.g. Tanenhaus et al. 1995, Kamide et al. 2003, Trueswell et al. 1999). In this paradigm, participants are presented with visual scenes either on the screen or directly as a layout of objects on a table. They hear, for example, sentences like (8) including objects while looking at a scene that either has one baby or two babies. It has been shown that adults do not really consider the PP "in the cradle" as a directional object when there are two babies in the visual scene (whereas children apparently do, see Trueswell et al. 1999).

(8) Put the baby in the cradle on the highchair.

Moreover, given a visually presented scene with only a few objects, the given context appears to show an immediate and dominating effect on disambiguation in general, but also on the anticipation of verbal arguments (Kamide et al. 2003). Visual scenes obviously present a very strong cue for

[2] We have to say that it is possible that prepositional phrases do not have the same properties as relative clauses in these constructions (we want to thank Bernhard Schröder for his comments on this problem). In the literature on parsimonious mental models, however, they have been treated alike.

a mental model of a situation described in a sentence or text. Within this paradigm it is often argued that a strong enough mental model is sufficient to eliminate any effect of syntactic processes. On the other hand, one may argue that the presentation of a visual scene is not only a very strong cue for a mental model, but also a very specific situation for language processing. It is a situation where listeners construct a strongly reduced discourse universe with only a few clearly defined objects. Most of the time, participants are allowed to scan the scene for a few seconds before the linguistic input starts. It is probably true that these experiments show that syntactic factors in ambiguity resolution can be eliminated in principle, but it is far from clear in how far this extends to all sorts of everyday language processing, where the discourse universe is far less constrained.

All in all, we can say that some kind of enriched discourse model certainly affects sentence processing, though sometimes only fairly late. However, lower level representations, i.e. syntactic and semantic representations, seem to play an important role as well, and maybe even more so in earlier stages of processing.

There is another more theoretical problem with the approach presented in this section. Research on referential parsimony, since Crain and Steedman, assumes that the complexity of a mental model depends on the number of entities it contains. In the following section, we will discuss whether this is a viable assumption.

5. The representation of numbers

The eliminative mental models approach presented in the last section implies a preference for parsimonious mental models. But when is a mental model parsimonious? A simple assumption would be that complexity depends on the number of entities to be represented in the model. If we assume the least complex model compatible with the current linguistic input, this preference for parsimonious representations can explain some interpretational preferences. However, is the number of entities in a mental model really relevant for its complexity? Or more general, how are numbers represented in a mental model (for a more detailed analysis of this problem, see Frazier 1999)? If a mental model is some kind of direct analogue or even a pictorial representation (e.g. Schnotz 2005), we have to assume that either ten houses are represented as ten single houses or they are represented as a bunch of houses (it would have to be specified, how that is going to work), but then their number should not be accessible anymore as soon as the text basis has decayed. Including some kind of symbols for bigger numbers in

mental model just means that we would be dealing with mixed representations, including propositional as well as non-propositional information.

If we compare the statements in 9 a-d, is the mental model to be assumed for 9b really more complex than the one to be assumed for 9a, or is the mental model for 9d more complex than the one for 9c? We are not aware of any psychological evidence suggesting that this is the case.[3]

(9) a. Few people attended the workshop.
b. Many people attended the workshop.
c. 10 people attended the workshop.
d. 50 people attended the workshop.

The problem is even more general than that. It has been shown that we have difficulties representing numbers in general. Consider the example by Sanford et al. (1994), (10) below.

(10) 70 % of my psychology class passed the statistics exam—Is that a lot?

In everyday life, we rarely reason with concrete numbers but mostly with vague categories like "a lot." How would numbers then be represented in a mental model? Consider the examples in (11). It is easily imaginable that the mental model for 11a contains two kids, even that the model for 11b contains three kids. But does the mental model for 11c really contain ten individual kids playing in the yard? Or is it just a bunch of kids roughly "tenish" running around?

(11) a. Two kids are playing in the yard.
b. Three kids are playing in the yard.
c. Ten kids are playing in the yard.

All humans are equipped with some ability for number representation from very early ages, but this ability only works with fairly small numbers (18 month old children can do basic arithmetic until up to three or four entities; Geary 1995). Even though cross-culturally mathematic competence differs considerably, this basic ability for number representation appears to be fairly stable. These small numbers may be what we directly represent in

[3] We would like to add that it is in fact very plausible that we construct a discourse model with only one girl, if only one girl is mentioned in the linguistic input. However, the reason for that is most probably not the increased complexity of a discourse model with two girls, but our adherence to the conversational maxime of quantity (be as informative as necessary; Grice 1975). Had more than one girl been relevant for the current universe of discourse, it would have been mentioned by the speaker.

our mental models. However, we do not represent individual entities for every tree in a forest of one thousand trees. Still we can use these numbers to make inferences. If we hear that six hundred of these trees were cut down for fire wood and four hundred others died from a disease, the forest is gone. So we obviously make use of the concrete numbers for abstract calculations, but not by the way of representing one thousand trees as single entities in a mental model. This calls for a level of semantic representation closer to the linguistic input, where quantifiers of various sorts are represented. [4]

The problem of number representation becomes even worse, when we have to take quantified or negated expressions as in (12) into account. [5]

(12) a. George drank two glasses of beer.
b. George drank at least two glasses of beer.
c. George drank only two glasses of beer.
d. George certainly only drank two and not three glasses of beer.
e. George did not drink a glass of beer.
f. George drank all glasses of beer he was offered, but nobody knew exactly how many.

These are of course the phenomena which are central topics of research in semantic theories (as in DRT, Kamp & Reyle 1993). Whereas these theories certainly would have to be expanded to formalize the spatial properties of mental models and their effects on language comprehension, it is very hard to see how quantifiers, negations and their interaction can be integrated into a mental model.

6. The mental model of a text as a basis for linking information: Syntactic and semantic processes in anaphor resolution

A central part of Garnham's definition of mental models in text comprehension was its utility for linking different parts of a text. A linguistic means to link information are anaphoric expressions like definite noun phrases (as "the woman" or "the girl" in 5, 6, 7 above) or pronouns. A very plausible assumption would be that finding the referent of an anaphor should mostly be based on the mental model of the current discourse. And in fact, this has been proposed by various authors.

[4] Similar arguments can be made for the representation of negation, where the linguistic input as well as the mental model of the situation are playing a role for the accessibility of discourse entities (Kaup & Zwaan 2003).

[5] We would like to thank Bernhard Schröder for these insights.

A seminal study by Gernsbacher (1991) looked at conceptual anaphora like (13 a-c). What is interesting here is that there is always a singular referent in the first sentence which is referred to by a plural pronoun in the second. In her experiments, Gernsbacher does not find any difficulty in processing the plural pronouns in sentences like these. She argues that participants consult their mental model of the situation, which is enriched by background knowledge, when they look for a referent of the pronoun. Since we know that there is more than one margarita in a bar, more than one plate in a kitchen and so forth, we can easily find the antecedent for the plural pronoun.

(13) a. I think I'll order a frozen margarita. I just love *them.*
b. I need a plate. Where do you keep *them*?
c. My mother's always bugging me to wear a dress. She thinks I look good in *them.*

However, the story may be a little more complicated than that. The plural pronoun here is only viable if the entity introduced in the first sentence allows for a non-specific, general, and/or collective reading. See for example (14a,b), where the singular pronoun in (14a) enforces the specific reading whereas the indefinite noun phrase "a Japanese woman" has to be interpreted as non-specific with the plural pronoun in (14b).

(14) Chris wants to marry a Japanese woman.
a. She is just the kind of woman he likes.
b. They are just the kind of women he likes.

What makes the so-called conceptual anaphor viable here is not the knowledge that there is more than one Japanese woman in the world, but that (14b) has a strong non-specific reading. What we need again, is a semantic level with constraints over specific and non-specific interpretations of noun-phrases that surely interacts with the mental model of the current discourse.

There is additional evidence that early processes in pronoun resolution are mostly based on lower level (mostly syntactic) principles. Sturt (2003) showed that only antecedents which are available according to syntactic binding constraints are considered in very early stages of processing, whereas all entities available in the mental model of the current discourse can be taken into account at later stages.

Another way to look at the kinds of processes involved in early anaphor resolution is the analysis of EEG-patterns while participants are listening to or reading sentences involving anaphoric violations. It has been proposed that different types of potential shifts directly following a linguistic event

(event related potentials, anaphor resolutions) correlate with different types of linguistic processes (see Friederici 1999, for a full model based on different components). A positive shift peaking at about 600 msec after the event (P600) is often found following syntactic violations (e.g. agreement violations as in "the girl sing;" Hagoort & Brown 1994). A negative shift peaking at about 400 msec (N400) has been established for semantic integration violations (e.g. as in "the boy kicked the milk," Kutas & Hillyard 1980). Hemforth & Frenck-Mestre (2005) present data from an anaphor resolution-study comparing sentences where the only possible antecedent within the sentence or text matched (a) or did not match (b) in gender with the pronoun in the second clause (15a,b, 16a,b). The experiments were done in French where gender is marked for singular as well as for plural pronouns. The antecedents were always humans and morphologically marked for gender, so that biological and grammatical gender matched. In one experiment, the antecedent appeared in the matrix clause of the same sentence, in a second experiment, it appeared in a separate sentence.

(15) Les bergères fluettes couraient vite quand (a) elles/ (b) ils ont rattrapé le troupeau.
The shepherdesses ran fast when they (a. fem, b. masc) recaptured the herd.

(16) Les bergères fluettes couraient vite. Enfin (a) elles/ (b) ils ont rattrapé le troupeau.
The shepherdesses ran fast. Finally they (a. fem, b. masc) recaptured the herd.

In both cases, Hemforth and Frenck-Mestre found a standard P600, suggesting that initial processes in anaphor resolution are actually more syntax than semantics based (see also Osterhout & Mobley 1995). Again, this evidence calls for several levels of representation, partly more syntactic or semantic in nature, partly more conceptual. Whether the N400 or any other EEG-component directly reflects the construction of mental models, is, however, surely an open question. To our knowledge, no clear-cut component has been associated with violations of mental model construction so far.

7. Conclusions

The general conclusions to be derived here are quite straightforward. Based on the evidence available from the literature and on our own experiments, we can clearly answer quite a few questions:

- Yes, conceptual representations play a central role in sentence processing.
- Yes, background knowledge plays a central role in sentence processing.
- Yes, with appropriate contexts and with appropriate tasks, we do construct conceptual representations with spatial and temporal properties.

But these answers cannot be taken as arguments for an eliminative approach trying to explain away linguistic levels of representation, because

- No, we cannot get rid of a discourse representation with strong linguistic links.
- No, on-line sentence processing is not (only) based on constructing and updating a mental model. Many other factors play a role as well.

Of course, this does not make the development of theories any easier. Single factor explanations are surely easier, but rarely correct. In order to arrive at a reasonable and correct theory of human language processing, we not only need a complex network of representations on various levels (linguistic as well as non-linguistic), we also need to explain how these networks interact. Moreover, we need to define which representations are built under which circumstances (in terms of a cost-benefit analysis). Jackendoff's theory of mental representations (Jackendoff 2002) may point into a useful direction here.

A major concern is the fact that there is no clear definition of what a mental model is when it comes to language processing. If it is a special kind of conceptual representation with particular constraints, then we will have to define these constraints, i.e we will have to define the conditions on which we would call something a mental model. Moreover, if mental models are to be taken as serious candidates for language comprehension, they have to deal with typical semantic issues like quantifiers, negations, and many more. Fortunately, the lack of a clear definition does not prevent us from trying to find more constraints on conceptual representations that will have to be taken into account.

References

Altmann, G. & Steedman, M. (1988), 'Interaction with context during human sentence processing', *Cognition* **30**, 198–238.

Bever, T. (1970), The influence of speech performance on linguistic structures, *in* G. F. d'Arcais & W. Levelt, eds, 'Advances in Psycholinguistics', North Holland Publishing Company Amsterdam, Amsterdam.

Britt, M., Perfetti, C., Garrod, S. & Rayner, K. (1992), 'Parsing in discourse: Context effects and their limits', *Journal of Memory and Language* **31**, 293–314.

Crain, S. & Steedman, M. (1985), On not being led up the garden-path: The use of context by the psychological parser, *in* D. Dowty, L. Karttunen & A. Zwicky,

eds, 'Natural Language Parsing', Cambridge University Press, Cambridge.

Desmet, T., De Baecke, C. & Brysbaert, M. (2002), 'The influence of referential discourse context on modifier attachment in Dutch', *Memory and Cognition* **30**(1), 150–157.

Frazier, L. (1987), Sentence processing: A tutorial review, *in* M. Coltheart, ed., 'Attention and Performance XII', 5th edn, Lawrence Erlbaum, Hillsdale, NJ.

Frazier, L. (1999), *On Sentence Interpretation*, Kluwer Academic Publishers, Dordrecht.

Frazier, L. & Fodor, J. (1978), 'The sausage machine: A two-stage parsing model', *Cognition* **6**, 291–325.

Friederici, A. D. (1999), The neurobiology of language comprehension, *in* A. D. Friederici, ed., 'Language Comprehension: A Biological Perspective', 2nd edn, Springer, Berlin, pp. 101–132.

Garnham, A. (1985), *Psycholinguistics: Central Topics*, Methuen, London.

Garrod, S. (forthcoming), Referential processing in monologue and dialogue with and without access to real world referents, *in* E. Gibson & N. Pearlmutter, eds, 'The Processing and Acquisition of Reference', MIT Press, Cambridge, MA.

Geary, D. C. (1995), 'Reflections of evolution and culture in children's cognition: Implications for mathematical development and instruction', *American Psychologist* **50**, 24–37.

Gernsbacher, M. A. (1991), 'Comprehending conceptual anaphors', *Language and Cognitive Processes* **6**, 81–105.

Glenberg, A., Meyer, M. & Lindem, K. (1987), 'Mental models contribute to foregrounding during text comprehension', *Journal of Memory and Language* **26**, 69–83.

Gorrel, M. (1995), *Syntax and Parsing*, Cambridge University Press, Cambridge.

Grice, H. (1975), Logic and conversation, *in* P. Cole & J. Morgan, eds, 'Syntax and Semantics 3: Speech Acts', Seminar Press, New York.

Hagoort, P. & Brown, C. (1994), Brain responses to lexical ambiguity resolution and parsing, *in* C. Clifton, L. Frazier & K. Rayner, eds, 'Perspectives on Sentence Processing', Lawrence Erlbaum, Hillsdale, NJ.

Heim, I. (1983), File change semantics and the familiarity theory of definiteness, *in* R. Bäuerle, C. Schwarze & A. von Stechow, eds, 'Meaning, Use and Interpretation of Language', DeGruyter, Berlin, pp. 164–189.

Hemforth, B. (1993), *Kognitives Parsing*, Infix Verlag, Sankt Augustin.

Hemforth, B. & Frenck-Mestre, C. (2005), Anaphor resolution within and across sentences: An ERP-study. Paper presented at CUNY, 21.March - 2.April 2005, Tucson, Arizona.

Jackendoff, R. (2002), *Foundations of Language*, Oxford University Press, Oxford.

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

Kamide, Y., Altmann, G. & Haywood, S. (2003), 'Prediction and thematic information in incremental sentence processing: Evidence from anticipatory eye movements', *Journal of Memory and Language* **49**, 133–156.

Kamp, H. (1981), A theory of truth and semantic representation, *in* Gronendijk, Janssen & Stokhof, eds, 'Formal Methods in the Study of Language, Part 1',

Mathematisch Centrum Press.

Kamp, H. & Reyle, U. (1993), *From Discourse to Logic*, Kluwer Academic Publishers.

Kaup, B. & Zwaan, R. (2003), 'Effects of negation and situational presence on the accessibility of text information', *Journal of Experimental Psychology: Learning, Memory and Cognition* **29**, 439–446.

Kelter, S., Kaup, B. & Claus, B. (2004), 'Representing a described sequence of events: A dynamic view of narrative comprehension', *Journal of Experimental Psychology: Learning, Memory and Cognition* **30**, 451–464.

Kimball, J. (1973), 'Seven principles of surface structure parsing in natural language', *Cognition* **2**, 15–47.

Konieczny, L. & Voelker, N. (2000), Referential biases in syntactic attachment, *in* B. Hemforth & L. Konieczny, eds, 'German Sentence Processing', pp. 135–157.

Kutas, M. & Hillyard, S. A. (1980), 'Reading senseless sentences: Brain potentials reflect semantic incongruity', *Science* **207**, 203–205.

Osterhout, L. & Mobley, L. (1995), 'Event-related brain potentials elicited by failure to agree', *Journal of memory and language* **34**, 739–773.

Sanford, A., Moxey, L. & Patterson, K. (1994), 'Psychological studies of quantifiers', *Journal of Semantics* **10**, 153–170.

Schnotz, W. (2005), 'Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverstehen aus der Sicht der Psychologie und der kognitiven Linguistik', *Jahrbuch der IDS* (in press).

Sturt, P. (2003), 'The time-course of the application of binding constraints in reference resolution', *Journal of Memory and Language* **48**(3), 542–562.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. (1995), 'Integration of visual and linguistic information in spoken language comprehension', *Science* **268**, 1632–1634.

Trueswell, J. C., Sekerina, I., Hill, N. M. & Logrip, M. L. (1999), 'The kindergarten-path effect: studying on-line sentence processing in young children', *Journal* **73**, 89–134.

Van Dijk, T. A. & Kintsch, W. (1983), *Strategies of Discourse Comprehension*, Academic Press, New York.

Vandierendonck, A., Dierckx, V. & Van der Beken, H. (2006), Interaction of knowledge and working memory in reasoning about relations, this volume, pp. 51–82.

Vosgerau, G. (2006), The perceptual nature of mental models, this volume, pp. 255–275.

Part IV

Philosophy of Mind

This Page is Intentionally Left Blank

Introduction: Philosophy of Mind

The final group of contributions considers the properties of mental models from a philosophical perspective. Two obvious distinctions have been touched upon all across this volume: the relation of mental models to mental images, on the one hand, and to perceptual representations, on the other. All articles in this part try to clarify one of these relations or both, and they do so through conceptual analysis and a new perspecitve at typical experimental data.

Johnson-Laird was the first to emphasize that mental models are not images. The distinction has led him to propose that Paivio's famous dual code hypothesis—mental representations fall into two classes: propositional representations and images—should be replaced by a triple code hypothesis: Cognition, in addition to propositional representations and images, requires a third kind of representation, namely models. In his 1983 book, Johnson-Laird proposes the set of images to be sub-set of models. In recent writings, he distinguishes both types by what they are meant to represent; mental models represent sets of situations while images represent particular situations. It is unclear whether the second proposal just specifies the first. Strictly, this would require to identify a singleton set, containing only one situation, with the situation, whence an image would represent a singleton set and thus would belong to the set of models. Psychologically, the identification, of course, is highly plausible but at any rate the issue calls for further clarification.

Some characteristics that could be criteria for distinguishing models and images easily come to mind. Images are rich in details, models are not. Images, but not models, are modality-specific, i.e. essentially related to specific channels of sensory input. Finally, models contain abstract elements, images don't. However, **Gottschling** argues that none of these criteria is really sharp. Her strategy is to look at a Marr-style hierarchical theory of vision and then identify images with perceptual representations in or above the middle of the hierarchy. The identification is argued for

by certain effects—such as optical illusions and visual neglects—that are connected with this level of visual perception and can be demonstrated in both vision and imagery. Now, at the intermediate and high level visual representations can already have degrees of richness in detail, can result from non-visual sensory input, and even can contain abstract elements—at least according to a certain version of the hierarchical theory. Hence, no distinction between images and models can be established. Gottschling's focus is Johnson-Laird's claim that models are representations of a more general type as opposed to images. The claim is supported by experimental results due to Knauff and Johnson-Laird that, according to these authors, show that imagery can get in the way of model generation and can impede reasoning. These results seem to show the more fundamental status of models for reasoning processes, given their interpretation is correct, but Gottschling aims to show that there is an alternative interpretation of the results that gives new plausibility to the image theory.

Vosgerau argues against Johnson-Laird's key thesis for distinguishing models and images: the presence of abstract elements in the former. According to Vosgerau, mental models owe their explanatory power to two features: They preserve the relevant structure of what they represent and they are natural representations in the sense that the relevant relations between the relevant elements of the represented situation are represented by perceptual relations (as opposed to abstract elements) in the model. Without the second feature, mental models could not gain explanatory advantage over other theories of reasoning, e.g. mental logics, as these latter also have the feature of structure preservation. It is thus essential to the theory of mental models to give the feature of naturalness more attention than Johnson-Laird does. Indeed, the feature of naturalness conflicts with the claimed presence of abstract elements in models. Weighing the explanatory power of the theory higher than its delineation from mental imagery, Vosgerau argues against abstract elements. He uses the example of negation, an abstract element that is frequently present in models according to Johnson-Laird, and proposes a different analysis.

Although Vosgerau ultimately remains silent about models and images, his argument raises doubt about Johnson-Laird's abstractness argument. In contrast to both Gottschling and Vosgerau, the paper by **Held** tries to re-establish the distinction by exploiting the idea that both types of representation have different functions and therefore represent sets of situations and single situations, respectively. The positive suggestion then simply is that a representation that has the function to represent a set of situations is a model while one that has the function to represent a single situation is an image—regardless of any content features.

It is sometimes argued that mental models, apart from their key role in reasoning, also play a role in perception. Some philosopher's, however, are

skeptical whether models have the same status in both cases, as they seem to be conscious in the former, but unconscious in the latter case (see, e.g., Rehkämper's contribution in Part 3 of this volume). In accordance with these misgivings, Gottschling's argument should not be understood as advocating that models or representations on par with them are perceptual representations. It is true that Gottschling argues for the idea that images can serve any cognitive purpose that models serve. Thus, there are no grounds for a distinction. Models and images are, in certain respects, functionally equivalent. Moreover, it is crucial to her argument that images are identified with intermediate or high-level perceptual representations. However, this identification proceeds via these representations' content, and Gottschling, at the same time, defends Kosslyn's idea that images and perceptual representations are functionally different: While the former result from top-down activation of information in memory, the latter work via bottom-up activation. In a similar vein, the paper by Held tries to establish a distinction between models and images, on the one hand, and perceptual representations, on the other. The key idea, again, is to work via the different functions of the different types of representation, but a further step is attempted by trying to answer the question how these different functions can be characterized informatively. The proposal is to introduce a new distinction among representations: objectual vs. non-objectual, and a criterion to decide when a mental representation is conscious or unconscious. As a result, it is argued that mental models do not play any role or have any function in perception.

This Page is Intentionally Left Blank

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Visual Imagery, Mental Models, and Reasoning

Verena Gottschling [1]

Universität Mainz [2]

Abstract
The focus of this paper is the relation between Steven Kosslyn's visual mental images and Johnson-Laird's mental models. Knauff et al. presented empirical evidence and a challenging argument for the hypothesis that in fact "visual imagery impedes reasoning." I argue that these results may look embarrassing for pictorialists, but closer inspection suggests that they are actually harmless. I argue that the presented evidence fails to show that imagery impedes reasoning. I present some objections to the explanation proposed by Knauff and Johnson-Laird by pointing out some terminological and conceptual problems. Afterwards, I sketch an alternative explanation, which is more pictorialist in spirit. In fact, even from the view of pictorialism, the results are not as surprising as they may seem at first blush. Finally, I claim that mental models and visual images are not as different as typically assumed.

1. Introduction

To determine the role of visual imagery in reasoning processes or, more precisely, in deductive reasoning, is a challenging task. Before this question can be answered, it is necessary to determine the relation between Steven

[1] I am grateful to Gerhard Strube and Carsten Held as well as to Gottfried Vosgerau for helpful comments on an earlier version of the paper.
[2] E-mail: gottschl@uni-mainz.de

Kosslyn's visual mental images and Johnson-Laird's mental models. This is the focus of my paper. Imagery is normally seen as an important cognitive capacity used for solving problems, and for learning and reasoning processes. That is what the so-called 'pictorialists,' including their leading researcher, Steven Kosslyn, maintain. Recently, Markus Knauff and Philip Johnson-Laird and colleagues argued that this picture is fundamentally misleading. Because mental images are much richer and more complex than mental models, mental models are more useful than mental images in many reasoning processes. In recent publications Knauff et al. (2000, 2002, 2003, 2006) and Knauff & Johnson-Laird (2002) presented empirical evidence and a challenging argument for the hypothesis that in fact "visual imagery impedes reasoning." If the 'imagery-impedes-reasoning-hypothesis' is correct, it calls for a significant change in the way we evaluate the imagery capacity. Therefore the imagery-impedes-reasoning-hypothesis raises several important issues regarding the relation between imagery and mental models.

These results may look embarrassing for pictorialists, but closer inspection suggests that they are actually harmless. I argue that the presented evidence fails to show that imagery impedes reasoning. I will also argue that a more specific claim—that conscious imagery impedes reasoning under special circumstances—is also problematic. I concentrate on two issues:

(i) In the hypothesis, the central term 'image' is used for a very special and very rich subclass of what 'images' are in pictorialism.

(ii) Moreover, an additional hidden premise is necessary for the argument and indispensable for the conclusion. This premise is not plausible and relies on strong assumptions.

Furthermore, there are alternative pictorialist explanations of the data available. I present some objections to the explanation proposed by Knauff and Johnson-Laird by pointing out some terminological and conceptual problems. Afterwards I sketch an alternative explanation, which is more pictorialist in spirit. In fact, even from the view of pictorialism the results are not as surprising as they may seem at first blush. Finally, I claim that mental models and visual images are not as different as typically assumed, and I will try to broker a reconciliation.

2. Mental models vs. visual images

Johnson-Laird originally assumed that mental models can take many forms and serve many purposes, and that visual images are a special case

of very rich mental models (Johnson-Laird 1983, 410, 446).[3] Later, he amended his thesis and maintained that visual images are a distinct form of representation and need distinct forms of processes. This is despite the fact that they often function like models and that both sorts of representations are much more closely related than the third kind of mental representation, propositional representations. He called this the 'triple-code' hypothesis.

In mental models, a structural isomorphism is to be found. Parts of the model correspond to the parts of what it represents. Johnson-Laird (1993, 16; 1996; 1998, 447) describes three characteristics of mental models:

(1) Each relevant part of the represented entity is represented by a corresponding token in a mental model.
(2) The properties of entities are represented by the properties of their tokens.
(3) Relations among entities are represented by relations among their tokens.

It is assumed that visual images as well as mental models can be used in reasoning processes. Furthermore, both kinds of representations have a structure that corresponds to the world.[4] However, the relations and properties between which this structural isomorphism exists differ. Therefore, mental models are not to be identified with visual images. (Johnson-Laird 1993, 160; 1998; Knauff & Johnson-Laird 2002)

Mental models are more abstract than images, they are basically spatial representations, whereas images are visual representations. The spatial information is represented on a scale or in a spatial array. In mental models, visual details such as color, texture, and form can be neglected. Furthermore, mental models are not restricted to a specific modality. It is possible that in mental models there is only a minimal degree of analogical structure, such as the use of separate elements to stand for different individuals. Moreover, in contrast to images, mental models can contain symbols: Even tokens representing abstract concepts like negation or quantifiers are allowed. But mental models can be used to generate visual images; 'mental model' is the more general notion and mental models "underlie" visual images. (Johnson-Laird 2001, 434; Knauff & Johnson-Laird 2002, 364)

In mental model theory, it is assumed that information from long-term memory is used to generate a mental model. In an additional step, subjects sometimes use the produced model, supplemented by additional information from long-term memory, to generate an image. Whereas the mental model is basically a spatial representation and can contain symbols, the image is richer, it contains visual information. For that reason, a model can

[3] Johnson-Laird used the term "rich." What is meant is a complex model.

[4] Or—if it is assumed that the world is not preorganized—to the conception of the world.

represent a set of alternative classes of situations; it cannot be visualized, in contrast to a visual image. This is the reason why mental models are a distinct form of mental representation.

According to the mental model theory of deduction, subjects construct a model, or a set of models, based on relevant background knowledge, the meaning of the premises, and perceptual information. They then formulate a conclusion by describing a relation in the model that was not explicitly asserted in the premises. Finally, they check whether there are any alternative models that are compatible with the premises but refute the conclusion. The conclusion is valid and necessary, if the conclusion holds in all models of the premises. [5]

One of the knotty problems with mental models is to determine precisely what they are. Mental models can neglect special aspects, but they do not have to. Johnson-Laird distinguishes physical and conceptual models. A model can be a "physical" model and consist of elements corresponding only to perceptible entities, in which case it may be realized as an image, either perceptual or imaginary. Thus, images are one of the major types of physical models (Johnson-Laird 1983, 422). Alternatively, it can contain elements corresponding to abstract notions; these are "conceptual" models. In case they do not neglect the form of the respective represented object, the question arises: What distinguishes mental models and visual images? Johnson-Laird characterizes visual images in contrast to models as representations of the perceptive aspects of a situation from an observer's point of view (Johnson-Laird 1996, 93). A visual image is a "vivid mental image that can contain concrete persons, objects, colors and forms and resembles real percepts" (Knauff et al. 2003, 567). It is also modality-specific. Knauff claims that parts of, relations in, or properties of the image represent persons, objects, colors and forms. Nonetheless, the mentioned similarity to percepts gives us a hint about how to determine the difference. For Knauff and Johnson-Laird, the experience of having an image is an essential part of visual imagery; visual images are conscious (Knauff & Johnson-Laird 2002, 364, Johnson-Laird 1998). Thus, we have good reasons to read them as claiming that the decision of whether an image or a mental model is used depends on introspection. If it 'feels like' perceiving—or at least similar—

[5] In cases of connectives like 'if,' 'and,' or 'or,' subjects construct a set of models in which each model represents a different possibility. The complete mental models describe exactly the cases where the compositional statement is true. The analogy to truth tables is obvious. Deductions that depend on quantifiers like 'all' or 'some' call for the construction of models containing sets of tokens in which each token represents an individual. Again, subjects generate as many mental models as possible and see whether the conclusion is true. If they fail to find a mental model in which the conclusion is false the composed statement is assumed true.

and is therefore modality-specific, it is an image, otherwise it is a mental model.

3. The empirical findings: Spatial and visual relations in deductive reasoning

Both Markus Knauff and colleagues and Knauff and Johnson-Laird have gained some surprising results from a number of simple reasoning tasks (Knauff & Johnson-Laird 2002). In simple deductive reasoning tasks (see below), they tested spatial versus visual relations. I will restrict myself to their basic ideas and findings. Examples of visual relations include cleaner/dirtier and fatter/thinner; examples of visual-spatial relations include above/below, front/back, etc. Finally, examples for control relations include better/worse and smarter/dumber. The reasoning tasks looked like the following: [6]

(1) The dog is thinner than the cat.
(2) The ape is fatter than the cat.
Does it follow?
(3) The dog is thinner than the ape?

The results were unexpected. People need longer to solve the visual relation tasks than to solve the spatial relation tasks. In addition, only in the case of the visual task is there activation to be found in areas of the visual system associated with the secondary visual cortex. However, there was no increased activation in the primary visual cortex (V1). In all tested relations, there was activation in the parietal cortex—the dorsal stream and higher order visual areas—which many scientists assume to be essential for spatial representation. Knauff and Johnson-Laird's conclusions were twofold; first, there is no support for the assumption that areas usually activated in visual imagery (i.e. the primary visual cortex and the ventral system) are activated during reasoning. Secondly, in the case of visual relations, activation in the secondary visual cortex correlates with longer reaction times (Knauff & Johnson-Laird 2002, 365f., 369ff.; Knauff et al. 2000, 3961; Knauff et al. 2003, 567). In my view, the pressing problem is to find both an explanation for the results and to address the implications of these results. This is the focus of the rest of my paper.

[6] I confine myself to experiment 1, see Knauff & Johnson-Laird (2002).

Relations	Correct responses	Mean response latencies in sec	Activated brain areas
visual	86 %	2.65	*secondary visual cortex (V2)* *parietal cortex*
visual-spatial	90 %	2.20	*parietal cortex*
control	92 %	2.38	*parietal cortex*

4. The imagery-impedes-reasoning-hypothesis

Knauff, Johnson-Laird, and colleagues argue that their findings are evidence for the imagery-impedes-reasoning-hypothesis (Knauff & Johnson-Laird 2002, 364). This hypothesis states that because images are more complex and contain more details than models, they impede reasoning in some situations. More exactly, what impedes reasoning in these cases is the phenomenal experience of having the image.

Moreover, they argue that their results show that reasoning is normally based on abstract mental models and not on images. They understand their results as being inconsistent with or, at least, pointing to a tension in Kosslyn's theory, where images are considered to be representations in a visual buffer, which is located in the primary visual cortex (Knauff et al. 2002, 203, 210; Knauff et al. 2003, 559f.). The reasoning process itself is not affected by the imaginability of the premises. In reasoning, people can even use elements that cannot be visualized. In these cases, the mental model contains abstract elements like symbols, negations, or even quantifiers. "Because a model can contain such abstract elements, it can represent a set of alternative classes of situation, and so it cannot be visualized." (Johnson-Laird 1996, 123) Because mental models are a distinct, more abstract form of mental representation, they were used in the cases which had faster response times. In other words, the fact that images and mental models differ regarding the mentioned characteristics explains the empirical findings. In the case of visual relations, people spontaneously use imagery that is not pertinent to the reasoning process itself. Relations that elicit visual images contain details that are irrelevant to an inference. These dispensable processed details impede the process of reasoning. Thus imagery impedes reasoning. But it is not only the representation of too many relations that impedes reasoning; rather, the phenomenal experience of the image "gets in the way of reasoning" (Knauff & Johnson-Laird 2002, 364). This may have serious consequences for the role of our imagery capacity in cognitive processes:

> [...] a theory that relied on visual imagery as the medium for reasoning would be implausible [...] Similarly, such a theory cannot readily explain why relations that are easy to envisage visually impeded reasoning. (Knauff & Johnson-Laird 2002, 371)

Before I raise some objections to challenge this conclusion, we must clarify some central assumptions in pictorial imagery theories.

5. Imagery theory

Kosslyn is the leading researcher of the pictorialist view. He claims that people use pictorial mental representations—visual images—to solve problems, get new information, and also for reasoning processes.

According to Kosslyn, mental visual imagery involves the activation of information processing mechanisms at all levels of the visual system. Images are patterns of activations in a medium of visual buffer having the properties of a coordinate space, a matrix. He states that imagery occurs in a functional buffer, which is also used in the early vision system and uses the same kinds of representation. The central idea in Kosslyn's theory is that while perception works by bottom-up activation, there is top-down activation of information from memory in imagery. Visual mental images are activations in the visual buffer, which are the result of these top-down activations, i.e. they are not caused by immediate sensory input. Thus, they are one form of short-term memory representations. A central feature of the visual buffer is the attention window. The function of the attention window is to select some configuration of activity in one region of the buffer for further processing. According to Kosslyn, this further processing is the same as that which would have been carried out in perception (Kosslyn 1994, 76f). Once a pattern of activity is evoked in the visual buffer, it is processed the same way regardless of whether the activation was invoked by sensory input or information from memory. Moreover, it includes analyses in both the dorsal and the ventral system. It is important to note that Kosslyn's theory is a hierarchical theory. Images are considered to fulfill representational functions only when an 'interpretive function' is applied to them and extracts their meaning.

Mental images can be very different; subjects can generate very general, low-resolution images, but can also generate very rich and specific images. Normally, subjects are not conscious of how they build up objects one part at a time when they form images (Kosslyn & Koenig 1992, 433). However, if a task requires solving a specific problem, an image can also be generated voluntarily. The image generation process is assumed to be sequential and starts with a global image. This image is strong because it has been acti-

vated frequently. According to Kosslyn, images fade after being generated in the visual buffer and require more-or-less constant processes for their maintenance. Therefore, the maintenance process is already necessary during image generation. In a second phase, the image can be enriched with additional parts and details, perhaps in more than one step. They become more specific, more detailed, organized into figure and ground, and they can have important and less important parts. These more complex images are closer to visual experience. The generation of images depends on which different sources of information are involved, which pathways, and which other concepts are activated. The image generation process is more complex in the case of multi-part images, because different stored perceptual units have to be integrated to form the image. Besides the generation process, there are various other processes; maintenance, transformation, and introspection. Thus, the maintenance process is important for all other processes. Images fade within an average duration of 250 ms. (Kosslyn 1994, 101). Effort is necessary to refresh them. The process of image maintenance involves not only storage processes but also active processes. We have to refresh images because otherwise they do not remain long enough to be used in imagery tasks, which are normally at least two seconds in duration. Thus, image maintenance is fundamental to all the other processes—both for generating more complex images with different parts and also for the transformation and introspection of images.

The implications for the possibility of unconscious imagery are obvious. Unconscious imagery is not excluded in general—but we should expect most cases of and especially the more complex cases of imagery to involve consciousness and concentration. Kosslyn accepts Baddeley and Logie's model of working memory (Kosslyn 1994, Denis & Kosslyn 1999). In his view, working memory relies on both short-term and long-term memory. In doing so, he is in the same boat as Knauff and Johnson-Laird.[7] According to Kosslyn, images are one form of short-term representations. Here, a perceptual structure (the visual buffer) is used to activate information in long-term memory.

There are at least two characterizations of imagery in the literature, which are unfortunately not consequently distinguished. Imagery is often characterized as seeing in the absence of the appropriate input (Kosslyn 1994, 74). In my view, there are two possible ways to understand this characterization. Imagery might be construed as the having of a vivid conscious

[7] Thus, all researchers assume that processes of visual perception (involving the activation of stored information in long-term memory) largely overlap with the processes of working memory. It is important to be aware that an alternative view would be to see working memory as a system that is functionally separate and independent from perception and from long-term memory processes.

experience of something not physically present. Or, the characterization might be understood as saying that imagery is the use of all (or at least parts of) the machinery used in visual perception. Knauff and Johnson-Laird seem to have the first characterization in mind. As I pointed out, it should be expected that most cases of imagery are conscious. Nonetheless, we have to distinguish both characterizations sharply from one another, even if both characterizations are conflated in the literature. In the second case, we are not talking about experiences but about the machinery used in perception. The second characterization is about similar representations in different levels of perceptual analyses; it is about similar processes and identical brain areas. However, it is not restricted to conscious imagery. According to this view, the use of all parts of the machinery used in visual perception, without an appropriate external input, might be sufficient for the conclusion that imagery takes place. Conscious awareness may, or may not, be involved. This allows for an important distinction. Regardless of which characterization we favor, conscious awareness during imagery might give us evidence that imagery takes place. If there is no conscious awareness of imagery, it does not necessarily follow that there is no imagery involved. Furthermore, if conscious awareness of imagery processes is to be found, this may be due to properties of the underlying processes and representations. Therefore, images are often accompanied by the experience of having the image—but conceptually, both are different.

6. Two readings of pictorialism

Although Kosslyn is only committed to the postulation of 'functional images' and a 'functional buffer,' he often postulates that images are pictorial in a stronger sense. He identifies the medium of the buffer with retinotopic visual areas and the higher level visual areas with higher processing stages. We need to systematically distinguish both pictorialist claims: Pictorialists can posit either 'only functional' images or 'really spatial' images.

On the first reading, it is stated that images function in a picture-like manner but are not "really" pictorial or spatial. Of course the challenging task is then to characterize the notion of functional images and space. As Pylyshyn puts it:

> The hard problem is to give substance to the notion of a functional space that does not reduce it to being either a summary of the data, with no explanatory mechanisms, or a model of a real literal space. (Pylyshyn 2002, 167)

In fact there is progress in this area: Michael Tye's AI-inspired account seems promising. The central idea is that implicit representation of dis-

tances occurs via the determination of neighboring parts of an entry in the matrix. In my view, this theory can be easily enhanced (Tye 1991, Rehkämper 1991, Gottschling 2003). I have dubbed this the '**CO**rrelated **RE**lations theory' or 'CORE-Theory' of imagery. There are several forms of pictorialism; CORE-Theory introduces basic constraints for any explanatory pictorialist account. Regarding the representation of spatial relations, a pictorialist is only committed to a functional representation of spatial relations, not real pictures in the intuitive sense. The visual buffer he is talking about can be a functional buffer, and spatial relations can be represented via other relations with the same number of places, as long as these have the same inherent constraints. However, they do not have to be represented by the identical relation.[8] On this functional reading, the answer to the question we have raised is obvious by definition. Even if the experience during visual imagery and perception is similar, the underlying representations may differ. Internal and external pictures do not have much in common; talking about internal pictures is a loose analogy.

The second reading, which is of greater interest for us, is stronger; images, as well as some perceptual representations, are identified via topographical mapping. On this account, these representations should be understood as activated or (in the case of images) reactivated patterns of activation in the visual cortex.[9] Note that this reading is not committed to a reductionist view of the mind-body problem. We are not committed to simply identifying images and their neural correlates by maintaining type-identity. Nor do we want to confuse and shift levels from the functional representational level to the level of neural properties. Ned Block (1987) distinguishes between *a priori* functionalism and the other available view, psychofunctionalism. Whereas *a priori* functionalism understands functional analysis as analysis of the meaning of mental terms, psychofunctionalism understands functional analysis as a scientific hypothesis. Psychofunctionalists claim that images as mental (phenomenal) states can be identified with functional roles. Empirical science is the tool we can use to correlate these states with special functional roles, because the functional components are anatomically distinct. Thus, the functional organization is mirrored by the organization of our nervous system. According to this view, mental states

[8] For a description and more detailed analysis of this issue, see Rehkämper (1991, 1995) and what he—following Steven Palmer—called a 'natural-n isomorphism.'

[9] In fact, this reading shows a relation between two important issues concerning mental visual imagery. The first issue is whether mental imagery involves some of the same representations normally used during perception or whether it involves only more abstract, 'post-perceptual' representations. The second issue is the question whether images have a special format, which is depictive or spatial. The two issues are in principle independent but in fact closely related, given the fact that topographical organization plays an important role at certain levels of visual regions.

play certain functional roles, and carry information. To learn more about these roles and their implementation, we should study the brain. It follows directly that topographical organization at certain levels of visual regions is insufficient for the claim that something is pictorial; this organization must be shown to play a role in information processes. If the relevant neighboring neurons are not connected or information can be shown to be processed inadequately, the spatial layout plays no functional role. In other words, what matters is the connectivity in the relevant areas, not the spatial layout itself.

In some recent work, Kosslyn seems to advocate a similar position:

> [...] roughly half of these visual areas [of macaque monkeys] are *retinotopically mapped.* That is, the neurons in the cortical area are organized to preserve the structure (roughly) of the retina. These areas represent information depictively in the most literal sense; there is no need to talk about an abstract functional space akin to that defined in an array in a computer. (Kosslyn 1994, 13)

It is important to be aware that the focus in the imagery debate is on spatial relations. The debate is more or less silent about other properties that are typically regarded as pictorial, such as texture, part/whole relations and color. In recent publications, Kosslyn has called Tye's hybrid account "reasonable." Tye states that properties such as color and texture are represented elsewhere in a more abstract format and are connected by pointers to specific parts of the depictive representation (Kosslyn et al. 2002, 200). Barsalou's (1999) proposal is very similar. Barsalou states that images are symbolic representations, which are nonetheless modality-specific, inasmuch as they consist of a subset of neural activity associated with the corresponding visual perception.

Thus images could be hybrid representations containing symbolic elements, but they are basically representations using a spatial layout (at least on a functional reading).

7. Visual perception, imagery, and mental models

7.1. Levels of visual perception

Vision is characterized as a multi-stage process. It is accomplished in three, roughly separable, and successive processing states, dubbed low-level, intermediate-level, and high-level vision. The traditional hierarchical view is that these stages encode different pieces of information about a

stimulus, and each then passes the information on to the next stage.[10] David Marr (1982) labeled the representations in low-level vision 'primal sketches.' These constitute retinotopically organized information, organized into blobs and edges. Low-level processing concerns momentary reflectance features and almost point-like areas. According to Marr's view, information about surfaces, depth, and shape is encoded in the viewer's perspective in intermediate-level vision. Intermediate-level processing analyzes a restricted set of features of longer intervals and larger surfaces, such as orientation, color, and distance; these representations in intermediate-level vision are called 2.5-D sketches. In contrast, high-level processing is concerned with object recognition, which is nonetheless vision-specific. The representations here, called 3-D models, are view-point invariant structural descriptions of objects, which can be matched with stored representations. The visual cortex consists of primary visual cortex and a number of other cortical areas that process different kinds of information.

Cognitive neuropsychologists distinguish successive states in vision corresponding to Marr's three levels together with three different levels of perceptual representations. These representations are called low-level, intermediate, and high-level representations. The functional description of these stages does not correspond exactly to Marr's account, but there are nevertheless widely accepted opinions about which regions of the brain correspond to which states in his account. Low-level vision is associated with the primary visual cortex (V1); it contains a retinotopically organized map of visual space, with adjacent locations in the retina corresponding to adjacent locations in the cortex. Intermediate vision is correlated with regions in the extrastriate cortex (V2-V4 and MT). Different layers of both V1 and V2 respond to different features (color, color-form motion, and dynamic-form). The most appropriate candidates for Marr's high-level vision and object-centered representations seem to be regions in the inferior-temporal cortex (IT) in the ventral system. Sometimes, areas in the prefrontal cortex (PFC), which receive highly processed visual information from IT, are also mentioned. Area V4 sends information into the inferior temporal cortex, where representations are abstracted away from details of a specific vantage point (such as lighting or location in the visual field). For our purpose, it is sufficient to understand these areas as the "perceptual front end" (Jackendoff 2002, 347).[11] The many feedback connections from high to low-level areas are assumed to mediate so-called recurrent processing, where low- and high-level information processing interact. This feedback from higher areas to V1 is assumed to play an important role in the generation of the surface

[10] Recently, this strict hierarchy is being questioned, but for our purposes it is sufficient to retain the traditional labels.

[11] Even if recently Marr's 3-D model is disputed in some vision research.

representation. The feedback is seen as the neural basis for segregating figure from ground, including motion, depth, and color. Therefore, the idea of informationally encapsulated levels in vision is a case of oversimplification (see Kosslyn 1994, 15f).

7.2. Images and levels of perceptual representation

The obvious question is then with which kind of perceptual representations images should be identified. If we look for the neural correlate of the visual buffer, the primary visual cortex (i.e. the most peripheral of the proposed levels) is in fact not the only available candidate. There are further candidates for retinotopical regions in visual areas. V1 is not the only region that is retinotopically mapped: V2 also contains topographically organized areas, and some other areas of the prestriate cortex contain such areas as well (V3, V3A, V4). Additionally, V2 sends efferent fibers to V1 and strongly affects the cells in area V1. In other words, the primary and secondary visual cortex are highly connected. If we are looking for the neural correlate of the buffer, we are looking for retinotopically mapped areas, and the spatial layout has to play a functional role. The reason is that this allows representations in the visual buffer to "contain an enormous amount of implicit information about spatial properties" (Kosslyn 1994, 86). However, it is not necessary to locate the buffer in the primary visual cortex. [12] If we systematically assess our options, we could state that images are perceptual representations (i) at a low level (V1); (ii) at an intermediate level (V2-4, MT); (iii) at a high level (IT); or (iv) at different levels. In fact, conscious perceptual images must be identified with intermediate level of perceptual representations to account for both recent empirical findings and philosophical considerations (Jackendoff 1987, Kosslyn & Thompson 2003, Kosslyn et al. 2002, Prinz 2000, Gottschling 2003, 2005).

> Images do not preserve the earliest visual representations. [...] Rather than being like "primal sketches" (in Marr's terminology), they are like 2.5-D sketches; they incorporate organized units. (Kosslyn et al. 2002, 198)

There are good reasons for this view: To begin with, many effects assumed to be located in the intermediate level—such as optical illusions, visual neglects—can be demonstrated in imagery as well as vision. But this is also compatible with the 'low-level' view, which assumes further analyses of the images in the buffer. More convincing evidence is that blind persons and especially persons born cortically blind with neglects in V1, are able to solve special tasks that seem to require imagery capacities. Marmor &

[12] For an overview of related empirical studies see Kosslyn & Thompson (2003) and Farah (2000).

Zaback (1976) showed that these persons are able to solve a variation of Roger Shepard's famous rotation experiments. Marmor and Zaback used drop-shaped sheets with left outs instead of letters. The participants were asked to decide whether it is possible to match two of these or not (without turning them around). The reaction time for the answers was linear, as in Shepard's classical rotation experiments. The primary visual cortex is destroyed in these cases. The usual reaction from pictorialists was that we have to distinguish visual and spatial representations in imagery. Blind persons are able to solve special imagery tasks that require only the representation of spatial relations by using higher levels of visual analysis. Thus, mental images should not simply be identified with perceptual representations in the visual buffer. Instead, there are two different kinds of images. Typically, we have top-down activation in visual imagery to the buffer; these *visual images* can be simpler, or more detailed and complex with two components. There is a short-term representation, which is pictorial in the literal sense. The medium of these representations is the visual buffer. The second component of an image is the information held in long-term memory, which is necessary to generate the short-term representation. After generating this representation, analysis occurs as in perception; that is form, color, and motion analysis and activation in the dorsal and ventral system. As well, there are pure spatial images, which we can use under special circumstances to solve imagery tasks. They are not represented in the visual buffer and do not require a spatial medium in the literal sense. In these cases, representation is much more abstract than in retinotopically organized areas in the visual cortex. In Kosslyn's view, the pure spatial images are normally part of the analysis of images in the visual buffer. However, they can also directly be generated from knowledge or from tactile input, as the examples of Marmor and Zaback show.

To summarize our results: First, these results show that low-level representations are not all that constitutes imagery—even if there is activation in the primary visual cortex in many cases. It is not only that one component of images is pictorial; information in images can be more abstract. Secondly, the term 'image' is an umbrella concept. The term is used ambiguously in the literature and is used for low-level representations as well as for pre-organized units like intermediate representations. Extending the use of 'image' to representations that do not have any functional spatial characteristics at all is a clear abandonment of the picture analogy. If no component of an image is functionally spatial (not to mention pictorial in a stringent sense), there is nothing left of the initial analogy.

Images are confined to a particular point of view. 3-D sketches are not. For that reason, high-level vision is not the main correlate of images but is nonetheless necessarily involved in imagery. There is strong empirical support for this claim in the famous debate about the possibility of rein-

terpreting images of ambiguous figures like the duck/rabbit (Reisberg & Chambers 1991, Mast & Kosslyn 2002). The stored category plays a causal role in the selection of contained information in an image, as well as for attention mechanisms. An image has to contain a descriptive element, a specification of properties such as orientation and figure/ground organization. Intermediate representations are essentially involved, but mediation from the higher-level plays a central role as well. The whole idea of interpreting and transforming images takes this for granted. If you look closely, it is even more complicated. Sometimes, an image is taken to be the conjunction of a quasi-pictorial component (Kosslyn 1981, 213) and a descriptive component stored in long-term memory. This means that only one part of an image is pictorial. Thus, images have two components: A short-term representation, which is 'quasi-pictorial,' and the descriptive information in long-term memory, which is used to generate the short-term memory representation. Therefore, it would be hasty to identify one component in Kosslyn's theory—the activation in the buffer (independent from the question of whether we locate it in primary or secondary cortex)—with the required image and neglect its embedding in the whole theory. While intermediate representations are necessary for imagery, this does not mean that intermediate representations alone constitute imagery or the experience of imagery. In fact, we need activation at both the intermediate and high levels.

Again, it seems that a simple identification of the neural correlates of images with activation in V1 or areas in extrastriate cortex (V2-V4 and MT) would be rash. It seems more appropriate to identify images with Marr's 2.5-D sketches. But this is not enough. We should not identify the correlates of images as intermediate representations alone. Images incorporating organized units provide us with the essential hint; the information has to be adapted by the information contained in high-level vision, and attention mechanisms also have to play an important role. In the meantime, it is empirically well-supported that there is little in visual scenes that is encoded directly. Therefore, we explicitly have to attend to the items in question by turning our attention to them (Henderson & Hollingworth 1999, O'Regan et al. 2000, O'Regan & Noë 2001). Thus, the feedback connections and knowledge about visual appearance, which is located in high-level vision, play an essential role. An interaction between high-level and intermediate representations in visual imagery is indispensable. This is in consonance with Kosslyn's hierarchical theory and the function of the attention window. Images are subordinated to descriptive representations. Indeed, substantial empirical evidence indicates that some high-level processes influence behaviors that are traditionally considered low-level or intermediate-level. This is also in accordance with Kosslyn's imagery account.

Pylyshyn claims that the close connection between images as short-term representations and corresponding knowledge in long-term memory, which is usually thought to be descriptive, is problematic. He argues that introducing conceptual complexity is the first step in the direction where "one gives the actual image less and less of an explanatory role" (Pylyshyn 2002, 178). This, however, should not imply that every proposal of this kind has *no* explanatory power at all. If including conceptual information in a theory of imagery has this consequence, then no hierarchical depictive theory of imagery is, in fact, possible. As far as I am aware, all proponents of hierarchical pictorialism admit that they need conceptual information and that high-level vision is essentially involved in imagery. The whole idea of generating an image in short-term memory from stored descriptive information to make implicit information available takes that for granted.

To summarize, visual images are not raw displays—as intermediate representations they are pre-organized. This is not to say that intermediate representations alone constitute imagery or the experience of imagery. In fact, we need activation at the intermediate and high levels. The image itself is an intermediate representation (a 2.5-D sketch)—a visual representation, but it is strongly bound to knowledge about visual appearances, and information about this lies in high-level vision. [13]

7.3. Mental models and levels of perceptual representation

What about mental models? I introduced the term in the sense of mental models theory. It is tempting to conclude from the claim that "models underlie images" that mental models are 3-D sketches. In these high-level representations, knowledge about visual appearance is encoded. We could describe them as hierarchical representations, as "prototypical instance[s] of a category" (Jackendoff 1992, 44). Alternatively, they could be described as image-schemas from which "a variety of images can be generated" (Jackendoff 2002, 347), and which encode possible shape variations of objects. When imaging occurs, the underlying mental model of the object is used to generate and update a new representation of its surface from a particular point of view. As Johnson-Laird puts it:

> [...] when you form an image, you must compute the projective relations from the model to the 2.5 D sketch: a model underlies an image. (Johnson-Laird 1983, 157)

[13] Jackendoff's term "spatial structure" makes use of Marr's 3-D sketch and understands Biedermann's geons as an extension. Jackendoff understands this structure as modality independent—in contrast to Marr, who sees it as a part of vision.

Although mental models have imagistic elements, they are not strictly visual. They are abstract and support visual object categorization and identification. There are image-schemas, abstract structures from which a variety of different images can be generated. In addition, a variety of percepts and images can be compared (or recompared). Information at this level is represented geometrically. Nonetheless, it is not restricted to a particular point of view. The represented information is also more abstract than experienced images and percepts. Note that we do not simply have a 3-dimensional object, we have a complex hierarchy of representations which include all parts of the object. For example, the 3-D representation of a human figure can be encoded in more detail; a representation of the arm consisting of the upper and lower arm, while the forearm is elaborated into arm plus hand part, which again is elaborated so as to include the five fingers etc. Thus, the 3-D representation encodes information about how objects are assembled out of parts.

The identification of high-level representations and mental models seems true to Johnson-Laird's intentions (Johnson-Laird 1998). But this strategy is not plain sailing. First, mental models are introduced as entities we use in reasoning processes and not only as the representations encoding the relevant information. On this reading, mental models are identified not with representations in working memory themselves but with representations of spatial structures, which are amodal; that is, they are not part of the perceptual systems any more (Jackendoff 1987, 2002). Second, they would be sets of hierarchical representations. However, this last objection can be quickly defused. Mental models are not sets of representations, but tokens from these sets are special situations. In the same way, the first objection is not a fatal one in its recent form. It shows only that mental models can have different functions and are insufficiently described. Nonetheless, there seems to be a derived form of this objection hidden here. Consider two points: (1) Images can be of different forms, but are closely connected with intermediate and high-level representations, and (2) models can take different forms as well and are to be identified with high-level representations. From this, it seems to follow that both kinds of representations are strongly connected and not independent forms of mental representations as stated in the triple-code hypothesis. In fact, we can expect to find activations at both levels and that both kinds of representations are involved in most reasoning tasks. In these cases, we need frequent reactivation from information stored in the 3-D representations. This implies that the conclusion that imagery impedes reasoning does not follow. For, as we have seen, the triple-code hypothesis is a necessary premise for that conclusion.

Let us take stock. We have seen that the pictorialist's view about the relation between spatial and visual representations is opposit to Johnson-Laird's position. The cases that Kosslyn regards as normal are excep-

tions for Knauff and Johnson-Laird and *vice versa* (see figure 1). Whereas Johnson-Laird identifies images with rich visual representations necessarily accompanied by some experience, the pictorialists' or, more precisely, Kosslyn's use of the term 'image' is extremely general. In fact the term 'image' is used for almost every short-term representation that represents spatial relations—regardless of how these relations are internally represented. For pictorialists, there is only one additional constraint: Inherent constraints of the represented relation also have to be in the representing relation. Kosslyn's images can be very simple or very rich. Color and texture can be

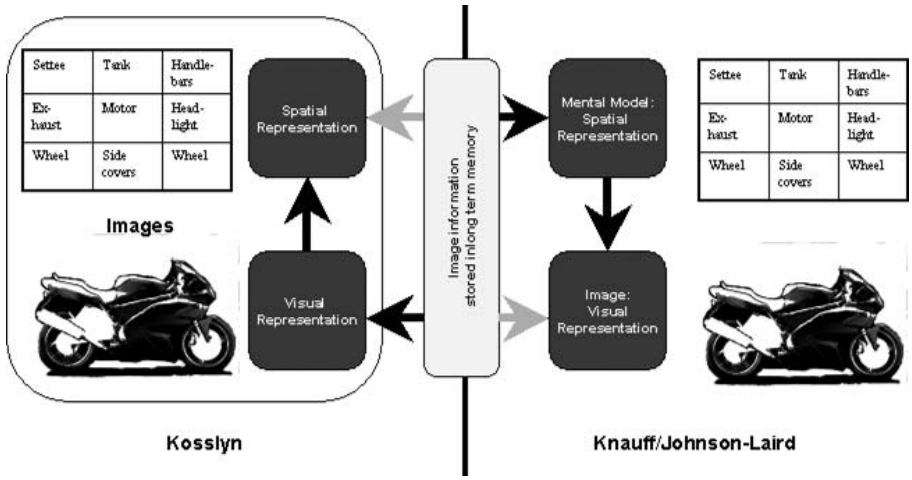


Fig. 1. The role of the "images" used in Knauff & Johnson-Laird's and Kosslyn's accounts

represented, but they are not necessarily elements of an image. Spatial relations, however, do not have to be represented via spatial relations. Let me explain this in more detail: For Kosslyn the essential properties of images are spatial relations. Other properties and relations we associate with images (such as color, texture, and part/whole relations) are optional. In some readings of pictorialism (see, for example, Michael Tye's interpretation of Kosslyn), these elements are represented via symbols or, more exactly, via pointers to symbolic entries in the cells of the array; so within images there are symbols and abstract elements.

If pictorialism in general is considered, the situation is even more complex (Gottschling 2003). At this point, an obvious objection is that Kosslyn's use of the term 'image' is too unspecific and too vague. Nonetheless, the lesson to learn here is that pictorialists like Kosslyn and Johnson-Laird use the term differently.

8. An alternative explanation

As I pointed out, the term 'image' in mental models theory is used differently from the way the term is used in imagery theories. This is particularly the case with Kosslyn. He uses the term in a much broader sense. In contrast to what Knauff and Johnson-Laird assume (Knauff & Johnson-Laird 2002, 370), images, as well as mental models, can neglect properties like color or texture. Additionally, images can be pure spatial representations that also contain symbols. Furthermore, images, particularly the simpler ones, can be unconscious. The argument that there must be some phenomenal experience associated with the image therefore cannot be run. We have seen that the term 'image' is vague and includes almost all visuo-spatial representations—but as we have also discovered, 'mental model' is an umbrella term in the same sense. This means that the vagueness objection applies equally to both parties in the debate. What about the results of the experiments? I have argued that the patterns of activation found in the experiments need not bother us. They are consistent with what we should expect. The fact that there is no activation in V1 is no threat for pictorialists. But the question of why subjects need more time to solve the visual relation tasks remains. According to Johnson-Laird and Knauff, the phenomenal experience during imagery impedes the reasoning process. I am not convinced by this argument. It is not decisive. The fact that we need more time to solve the visual relation task neither shows that we use images only in these cases nor that it has to do with conscious imagery. Furthermore, as I pointed out, to argue in terms of the experience that comes along with imagery does not seem a good strategy. For an interpretation of the delayed reaction times in the case of visual relations, we must look elsewhere. But even if this explanation of what impedes reasoning is misleading, what about a more compelling interpretation of the results that preserves the Knauff and Johnson-Laird's core idea? Perhaps in the case of visual relations too many unnecessary relations and properties are represented—maybe that is it what impedes the reasoning process. Such an explanation would be reconcilable with pictorialism.

The central question now is how spatial relations are internally represented in working memory. Of course, 2-place visual relations can be coded in spatial relations as well. But do they have to be? I have argued elsewhere (Gottschling 2003) that there are several possible ways in which Kosslyn's theory—and pictorialism in general—can be interpreted. But the only thing that is required is that spatial relations are represented via relations with two elements which have the same inherent constraints. There are two strategies available, both of which are consistent with pictorialism. First, we could assume that visual relations are internally encoded as visual rela-

tions, and that the participants were tempted to use more complex mental images than were necessary. If that were the case, the generation and maintenance processes would be more complex as well, which would slow down the process.

Secondly, we can turn to Kosslyn for another obvious explanation; subjects need less time to represent spatial relations as spatial relations than they need to represent other 2-place relations as spatial relations. It is important to be aware that the deductions we are interested in are very simple; only two contrasting relations are important. According to this explanation, what slows down the process is the coding in two other contrasting relations and not the coding of too many properties and relations not necessary for the deduction.

Thus, there are two possibilities: The process is slowed down either because too many details are represented, or because other 2-place predicates need to be coded. Note that, in Kosslyn's terms, images are used in all these cases. The first possibility is not threatening for pictorialists. In all cases we use images for the deductions, but in the case of pure visual relations we are, for some reason, tempted to use images that are too complex. What about the second possibility? Knauff and Johnson-Laird rule out this possibility from the beginning, but it is not clear why. The central question is which representing relation is used to code the represented relation 'is-dirtier-than.' One might use the identical relation [14] or another 2-place predicate with the same inherent constraints. Transferred to our example, 'is-left-of' might represent 'is-dirtier-than' or 'is-smarter-than.' Why should a spatial relation not be used to represent a visual relation like the 'is-dirtier-than' relation? Knauff and Johnson-Laird presuppose that this relation is internally represented in a vivid visual picture very similar to an external picture. But that confuses the phenomenal experience of having an image and the representation itself. The phenomenal experience caused by some property is not identical to the representation of that property. The way a property is experienced during imagery (the spatial relations, colors we experience, etc.) does not necessarily match with the way properties are represented in the underlying image. To conclude that a red experience is represented internally as red is not valid. The conclusion that our visual experience that 'a-is-left-of-b' is represented as 'is-left-of' in images is not valid either. For the same reason, we can look inside the brain and we will not find pictures or sentences. However, that does not imply that there are no pictorial or descriptive representations. To understand what a representation represents and how it represents we need to know how the representational system works. Both pictorialists and mental model theo-

[14] Thus, relations of the represented object are preserved by their physically equivalent relations of the representing object; there is a concrete first-order isomorphism.

rists should have little sympathy for the view that spatial properties are represented by using the identical relations. We should be careful not to confuse the phenomenal experience of using images with the image itself. There is no reason in principle why the visual relations used in the reasoning tasks cannot be represented via different visual relations, namely spatial relations.

In the case of mental models, Knauff and Johnson-Laird allow that the representing relations may differ from the represented relation. They assume that 'is-dirtier-than' can be represented in the same way as 'is-right-of;' therefore, spatial relations can encode non-spatial relations that are relevant for the deduction. However, they do not consider this option in the case of images. They need an argument why this possibility is excluded in the case of visual images because this is a necessary premise for their conclusion. An argument for this premise is in urgent need because pictorialist theories actually allow for more abstract coding. Moreover, to conclude that imagery impedes reasoning, it needs to be shown that pictorialism is committed to this premise. Thus, without further assumptions, the conclusion that imagery impedes reasoning is not warranted.

To summarize, there are two alternatives pictorialists can avail themselves of to answer the question of why people need more time to solve the visual relation tasks: (1) They are tempted to use more complex visual images than is necessary; or (2) they have to code the represented visual relation in another 2-place predicate. In the second case, we could expect that this kind of coding requires more time than is required for coding visual relations. On the pictorialist account, images are used in both cases.

9. Summary

I have argued that the evidence for the imagery-impedes-reasoning-hypothesis is unsatisfactory because there are alternative pictorialist explanations available. But the situation is even worse: If Knauff and Johnson-Laird's claim is merely that in some cases the coding of non-spatial relations via coding in different relations is slowing down the process, ramifications are harmless. So if this is what is meant by the claim that imagery impedes reasoning, they will not have any disagreement with pictorialists. Imagery does not impede reasoning in general, and Knauff and Johnson-Laird have attacked a straw man. But what about the weaker version of their thesis that I discussed? Does conscious imagery impede reasoning under special circumstances? Even that seems problematic. I have shown that Kosslyn's theory of imagery includes Knauff and Johnson-Laird's pure abstract representation of spatial relations on an axis (one dimensional) or

in an array (two dimensions). I have also shown that the conclusion to their argument is spurious, because their notion of 'image' is not really what images are. Images should not be identified with patterns of activations in the primary visual cortex (V1). Rather, it seems appropriate to identify images with intermediate representations (Marr's 2.5-D sketches) that are embedded in a highly connected system. Conscious perceptual images should be identified with intermediate level representations. That certainly weakens the intuitive idea of "pictures in the head," an idea that many participants in the debate are prone to use. But it is not threatening for the perception imagery analogy. As a result, the findings from Knauff, Johnson-Laird, and colleagues do not threaten pictorialists. The experimental results are inconclusive for their hypothesis. It is not only that alternative explanations of the data are available; to back up the presented argument they have to deliver two additional links: First, in order to argue for the imagery-impedes-reasoning-hypothesis, an additional premise is necessary. This would be the premise that the represented relation must be represented by the identical relation in case of images but not in case of mental models. This seems extremely difficult to argue for. Secondly, images, the underlying representations during imagery, have to be distinguished from the conscious experience of the imagery process. We must be cautious not to conceptually confuse the image and the phenomenal experience of having the image, even if the most interesting cases of imagery involve consciousness. To argue that the phenomenal experience during imagery gets in the way of reasoning can be dispensed with in favor of the imagery-impedes-reasoning-hypothesis. But whether or not the hypothesis can be proven, Knauff and Johnson-Laird are probably right to claim that "a theory that *relied* on visual imagery as the medium for reasoning would be implausible" (Knauff & Johnson-Laird 2002, 371; my italics). Why that? This clearly attacks a straw man. Visual imagery has the role of being a helpful aid we can sometimes use in the reasoning processes. To my knowledge, no pictorialist has ever maintained that imagery is the only medium of reasoning. However, the imagery-impedes-reasoning-hypothesis is questionable. As I have shown, pictorialists can easily find explanations why deductions with visual relations are more time consuming.

Leaving aside the hypothesis and its possible consequences, there are interesting results regarding imagery and the relation between mental models and visual images, for they highlight widespread misunderstandings of what images are. Images are not to be identified with the conscious experience of having them. Also, experienced properties of images are not necessarily coded in the representations by using the identical relations we experience. Moreover, if imagery theorists identify images represented in the buffer with perceptual representations and especially retinotopically mapped visual areas as the medium of the buffer, they are not compelled to identify

this area with V1. Rather, 'image' is an umbrella term that covers different kinds of internal spatial representations. Therefore, Kosslyn's images just include Johnson-Laird's mental models; or, to put it more carefully, both terms are currently used in an inflationary way. On closer analysis, Knauff and Johnson-Laird's view reveals a surprising kinship with Kosslyn's view.

References

Barsalou, L. (1999), 'Perceptual symbol systems', *Behavioral and Brain Sciences* **22**(4), 577–660.

Block, N. (1987), 'Troubles with functionalism', *Minnesota Studies in the Philosophy of Science* **9**, 261–325.

Denis, M. & Kosslyn, S. M. (1999), 'Scanning visual mental images: A window on the mind', *Cahiers de Psychologie Cognitive / Current Psychology of Cognition* **18**(4), 409–465.

Farah, M. J. (2000), The neural bases of mental imagery, *in* M. S. Gazzaniga, ed., 'The New Cognitive Neurosciences', 2nd edn, MIT Press, Cambridge, MA, pp. 965–974.

Gottschling, V. (2003), *Bilder im Geiste: Die Imagery-Debatte*, mentis, Paderborn.

Gottschling, V. (2005), Mental picture: Pictorial? Perceptual?, *in* K. Sachs-Hombach, ed., 'Bildwissenschaft zwischen Reflexion und Anwendung', Herbert von Halem Verlag, Köln.

Henderson, J. M. & Hollingworth, A. (1999), 'The role of fixation position in detecting scene changes across saccades', *Psychological Science* **10**(5), 438–43.

Jackendoff, R. (1987), *Consciousness and the Computational Mind*, MIT Press, Cambridge, MA.

Jackendoff, R. (1992), *Languages of the Mind*, Bradford, MIT Press, Cambridge, MA.

Jackendoff, R. (2002), *Foundations of Language*, Oxford University Press, Oxford.

Johnson-Laird, P. (1983), *Mental Models*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P. N. (1993), *Human and Machine Thinking*, Erlbaum, Hillsdale, NJ.

Johnson-Laird, P. N. (1996), Images, models, and propositional representations, *in* M. de Vega, M. J. Intons-Peterson, P. Johnson-Laird, M. Denis & M. Marshar, eds, 'Models of Visuospatial Cognition', Oxford University Press, New York, pp. 90–127.

Johnson-Laird, P. N. (1998), Imagery, visualization, and thinking, *in* J. Hochberg, ed., 'Perception and Cognition at Century's End', Academic Press, San Diego, CA, pp. 441–467.

Johnson-Laird, P. N. (2001), 'Mental models and deduction', *Trends in Cognitive Science* **5**(10), 434–442.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991), *Deduction*, Erlbaum, Hillsdale, NJ.

Knauff, M. (2006), A three-stage theory of relational reasoning with mental models and visual images, this volume, pp. 127–152.

Knauff, M., Fangmeier, T., Ruff, C. C. & Johnson-Laird, P. N. (2003), 'Reasoning, models, and images: Behavioral measures and cortical activity', *Journal of Cognitive Neuroscience* **15**(4), 559–573.

Knauff, M. & Johnson-Laird, P. N. (2002), 'Visual imagery can impede reasoning', *Memory & Cognition* **30**, 363–371.

Knauff, M., Kassubek, J., Mulack, T. & Greenlee, M. (2000), 'Cortical activation evoked by visual mental imagery as measured by functional MRI', *NeuroReport* **18**, 3957–3962.

Knauff, M., Mulack, T., Kassubek, J., Salih, H. R. & Greenlee, M. W. (2002), 'Spatial imagery in deductive reasoning: A functional MRI study', *Cognitive Brain Research* **13**, 203–212.

Kosslyn, S. M. (1981), The medium and message in mental imagery: A theory, *in* N. Block, ed., 'Imagery', MIT-Press, Cambridge, MA, pp. 207–244.

Kosslyn, S. M. (1994), *Image and Brain: The Resolution of the Imagery Debate*, MIT Press, Cambridge, MA.

Kosslyn, S. M. & Koenig, O. (1992), *Wet Mind: The New Cognitive Neuroscience*, The Free Press, New York.

Kosslyn, S. M. & Thompson, W. L. (2003), 'When is early visual cortex activated during visual mental imagery?', *Psychological Bulletin* **129**(5), 723–746.

Kosslyn, S. M., Thompson, W. L. & Gattis, G. (2002), 'Mental imagery doesn't work like that', *Behavioral and Brain Sciences* **25**(2), 198–200.

Lewis, D. (1986), *On the Plurality of Worlds*, Blackwell Publishers, Oxford.

Lycan, W. (1996), *Consciousness and Experience*, MIT Press, Cambridge, MA.

Marmor, G. S. & Zaback, L. A. (1976), 'Mental rotation by the blind: Does mental rotation depend on visual imagery?', *Journal of Experimental Psychology: Human Perception and Performance* **2**, 515–521.

Marr, D. (1982), *Vision*, W. H. Freeman, New York.

Mast, F. W. & Kosslyn, S. M. (2002), 'Visual mental images can be ambiguous: Insights from individual differences in spatial transformation abilities', *Cognition* **86**, 57–700.

O'Regan, J. K., Deubel, H., Clark, J. J. & Rensink, R. A. (2000), 'Picture changes during blinks: Looking without seeing and seeing without looking', *Visual Cognition* **7**, 191–212.

O'Regan, J. K. & Noë, A. (2001), 'A sensorimotor account of vision and visual consciousness', *Behavioral and Brain Sciences* **24**(5), 939–73.

Palmer, S. (1978), Fundamental aspects of cognitive representation, *in* E. Rosch & B. L. Lloyd, eds, 'Cognition and categorization', Erlbaum, Hillsdale, NJ, pp. 259–302.

Prinz, J. J. (2000), 'A neurofunctional theory of visual consciousness', *Consciousness and Cognition* **9**, 243–259.

Pylyshyn, Z. (2002), 'Mental imagery: In search of a theory', *Behavioral and Brain Sciences* **25**(2), 157–238.

Rehkämper, K. (1991), Sind mentale Bilder bildhaft?—Eine Frage zwischen Philosophie und Wissenschaft, PhD thesis, University of Hamburg, Germany.

Rehkämper, K. (1995), Analoge Repräsentationen, *in* K. Sachs-Hombach, ed., 'Bilder im Geiste. Zur kognitiven und erkenntnistheoretischen Funktion piktorialer Repräsentationen', Rodopi, Amsterdam, pp. 63–106.

Reisberg, D. & Chambers, D. (1991), 'Neither pictures nor propositions: What can we learn from a mental image?', *Canadian Journal of Psychology* **45**, 336–352.

Rosenthal, D. (1997), A theory of consciousness, *in* N. Block, O. Flanagan & G. Güzeldere, eds, 'The Nature of Consciousness', MIT Press, Cambridge, MA, pp. 729–753.

Seguin, E. G. (1986), 'A contribution to the pathology of hemianopsis of central origin', *Journal of Nervous and Mental Diseases* **13**, 1–38.

Tye, M. (1991), *The imagery debate*, MIT Press, Cambridge, MA.

This Page is Intentionally Left Blank

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

Mental Models as Objectual Representations

Carsten Held

Department of Philosophy, Universität Erfurt [1]

Abstract
In the debates of cognitive psychologists about the nature and cognitive role of mental models, the demarcation of such models and other types of representation plays a critical role. The purpose of this paper is to make the differences between kinds of mental representation sufficiently precise. Initially, I will isolate a distinction, largely overlooked in philosophical accounts of mental representation, but crucial for an understanding of the conception of a mental model—the distinction between objectual and non-objectual representations. I will draw further support for the distinction from two different attitudes involving a propositional representation a cognizer can have. As a result, I will criticize the contention that perceptual information processing requires the use of mental models. Finally, I will show that typical empirical results, interpreted in terms of mental models theory, also support the objectual/non-objectual distinction. Vice versa, the distinction can be utilized to better interpret these results.

In the debates of cognitive psychologists about the nature and cognitive role of mental models, delineating such models from other types of mental representation plays a critical role. The well-known debate about the function of imagery in human thinking may be cast in the question: Is *pictorial representation* really necessary to explain typical cognitive capacities and achievements of humans, or is it a mere epiphenomenon, playing no explanatory role in comparison with *propositional representation*? To appreciate what is at issue in this dispute, we require an understanding of

[1] E-mail: carsten.held@uni-erfurt.de

the difference between the propositional and the pictorial format. In a very similar debate, the cognitive role of mental models has been controversially discussed in a way that may be captured in the following questions: What does the *mental models* conception contribute to explaining typical cognitive abilities and achievements? Isn't much the same explanatory work done by the concept of *image-like representations*? Again, when we try to answer the question, a tacit understanding of different types of mental representation, mental models as opposed to images, is in play.

My purpose here is to make the differences between these kinds of mental representation sufficiently precise. At the outset, I will (1) isolate a distinction, well-known but seldom made explicit in psychological as well as philosophical accounts of mental representation—the distinction between transparent and opaque representations. I will argue that, crucial as the distinction is, the transparency-opacity metaphor is better replaced by an opposition of what I call objectual and non-objectual representations. I will (2) cross-classify this distinction with the one of pictorial and symbolic representations and will (3) try to establish a clear distinction of models and images. As a result mainly of the initial reflections (1), I will (4) criticize an often-heard contention in the mental models discussion, i.e. that perceptual information processing requires the use of mental models. Finally, I will (5) show that some empirical results exploited by Johnson-Laird to underpin his theory of the ubiquitous presence of mental models in human thinking support the objectual/non-objectual distinction. Thus, the distinction can, *vice versa*, be utilized to better interpret these results.

1. Objectual and non-objectual representations

The concept of mental representation inherits one characteristic from the general notion of representation that it will be helpful to explicate for what follows. Namely, a representation is an entity that *has*, for a representer, the *function* to represent something, the *representandum*.[2] The function terminology brings about some conceptual obligations. If some X is said to have the function to realize a goal Y, this is equivalent to saying that it is a means for the realization of an end Y. This follows from the fact that, given we can equate goals and ends, there is a total exchangeability of 'is a means for realizing' and 'has the function to realize' *salva significatione*. Now, the means-end terminology has logical implications. Namely, X is a necessary but not sufficient means for Y, i.e. the existence of Y entails the one of a means like X, but not *vice versa*. This logical feature trans-

[2] As one of many examples from the literature see Denis (1991*a*, 7–9)

lates back into function terminology: X has, but does not necessarily fulfill, the function to realize Y, and if Y is realized this entails the existence of something fulfilling the funtion that X has. Consider, for illustration, an example of visual perception. Suppose that a certain visual representation of mine has the function for me to make me see the Eiffel tower. If this function is fulfilled the Eiffel tower exists as my visual *representandum.* If it is left open whether or not the function is fulfilled, it is not left open whether or not the Eiffel tower exists, but whether or not it exists *as* my *representandum.* Now, suppose that it does so exist. Then this entails the existence of an appropriate visual representation. In general, the existence of the Eiffel tower as a *representandum* entails the one of a representation that has been sufficient to make it a *representandum.*

Against this foil, I wish to spell out a distinction among representations implicitly present in the philosophy and psychology of cognition since long. The best illustration, again, is perception. Philosophers have emphasized that the representations involved in the process of perceiving everyday objects have a crucial property: They are transparent, i.e. "we normally 'look' right through them." [3] The visual metaphor is meant to convey that perception involves representations, but their status in the perceptual process is different from the perceived objects' one. They have a function in the process but are not seen. This is not just an obvious triviality but also follows from the above remarks about representation, as it is the objects themselves that are perceived and perceiving something involves representing it so that the perceived objects are the *representanda.* The representations have, within the activity of representing the objects of perception, the function of making them *representanda* and as such they are ontologically independent of the latter. Their function is described as that they are 'looked through' or are—and here one metaphor replaces another—*transparent.* Good examples of what the metaphor intends are indeed found in the psychological literature on vision. When psychologists point out, e.g., that visual perception of three-dimensional objects involves particular "visual representations that precede the recovery of depth information" (Spelke et al. 1995, 309), they imply that these visual representations have a certain function in the recovery process, but not that they are themselves seen at any stage of it. Similarly, in Marr's famous theory of vision a "raw sketch" that is "2 1/2 dimensional" plays a key role (Marr 1982, 42), but, at the same time, Marr insists that we ultimately see the visual object *three-dimensionally.* This implies that the sketch performs a certain function in the visual process, not that it is itself seen (for that would imply that we simultaneously see

[3] van Gulick (1988, 178); the notion is explicitly introduced, under the title "diaphanousness" by G.E. Moore in 1903, but the phenomenon itself is aptly described already by Thomas Reid in 1764 (see Kind 2003).

two things of different dimensionality in the place of one).[4] In both cases, a difference is presupposed between an entity having a function in the process of seeing and the entity being seen. This difference arises from the very attribution of functions to representations and is nothing but the said ontological independence. The function concept, moreover, affords a way to cash out the transparency metaphor: A representation's transparency is its having the function of making something a *representandum*.

Consider, by contrast with the vision process, the kind of representation involved in processes of mental imagery. A typical case is a task, described by Kosslyn (1995, 268), to decide from memory what shape a beagle's ears are. The subject here must reproduce, from memory, an image of a typical beagle, thus, must produce a certain representation and then inspect this image.[5] Kosslyn initially likens the whole process to vision by saying that "visual mental imagery is 'seeing' in the absence of the appropriate immediate sensory input" (1995, 267). This description is inaccurate, to say the least. In contrast with vision and the representation there involved, in imagery the image is not a transparent representation. It is neither looked through within a process of seeing something else, i.e. a visual object, nor does it have a function within such a process, simply because the process is not one of seeing and no visual object is involved. Of course, the crucial difference with vision is not the mere absence of a visual object as *representandum*, but the lack of any intention to achieve one. In order to decide what shape a beagle's ears are, a cognizer must produce an image from memory and inspect this very image. Hence, it is the image itself she must attend to in the task and it is her immediate object, as is the visual object in visual perception. The cognizer is intentionally directed toward a representation, not, by means of one, toward a *representandum*. Returning to the earlier metaphor, it would be mistaken to call the image involved in the imagery task a *transparent* representation but appropriate to call it an *opaque* one.

To be sure, the image does have a function in the imagery task just as the perceptual representation has one in a perception task, but it is a different one. The beagle-image has a function for deciding from memory what shape a beagle's ears are and this task consists of three components: (a) producing the representation from memory, (b) inspecting it, and (c) conceptualizing the shape read off during (b). So, producing the image is only one constituent of the whole task. The image has a function in the process

[4] Is Marr himself aware of the distinction? Some of his formulations, e.g., about inferences from these sketches, suggest that he mistakes them as non-transparent, but there are others where he seems to suggest just the opposite (see his remark on the expression 'sketch,' (1982, 277).

[5] See Kosslyn (1995, 268-9) and his (1994, ch.s 9 and 10) for more on generation and inspection of images.

of performing the whole task, and it does so via (b) and (c), not (a). It has no function in (a), the process of its own production, because this would mean to confuse, with respect to one mental activity, its means and end. In a perception task, on the other hand, processing of sense-data is required, which results, at some step, in a visual representation. So a perceptual task, likewise, requires the production of a representation that cannot be a means during its own production or have a function in it. The difference comes only at a later stage (step (b) in the imagery task), where the subsequent constituent is described as the inspection of the representation produced on the one hand and the 'looking through' one on the other. After having been produced, both representations have their respective functions within different tasks. The transparent-opaque metaphor means to capture just these different functions. Since they are different and the proposal for transparency was to identify it with a representation's having the function of making something a *representandum*, its opacity can, at least, not be *this* function.

In order to pin down the latter function, another look at Kosslyn's description of the task is helpful. He points out that people reporting their own use of mental imagery in fact say that they "see" the image itself. They report, thus, being intentionally directed toward the image as if it were a visual object. Kosslyn himself accordingly speaks of "the imaged object" or "the object in the image" (Kosslyn 1995, 269) when the context clarifies that he does not mean an object distinct from the image but the image itself. The image itself *is* the cognizer's object. Of course, 'seeing' an image is different from seeing an object, but the point of comparison is that in both cases the cognizer is directed toward the entity in question—either toward the image she inspects or toward the visual object. As the expression 'being directed toward' invites misconceptions, it is best to compare the inspection of a mental image to a case of vision where the visual object is *inspected.*[6] In both these cases, the object and the image, respectively, have the same trivial function, i.e. to serve *as* objects of inspection. The cognizer has to have them *as* her object. This insight gives us both a clear translation proposal for the transparency-opacity metaphor and a criterion when one or the other case is realized. The criterion is this: If a cognizer reports a result of an inspection task involving a visual object, a minimally appropriate report will have to contain a singular term referring to the object. Likewise, if a cognizer reports a result of an imagery task, like Kosslyn's, a minimally appropriate report will have to contain a sin-

[6] Contrast the case with visual *detection* that might be reported by a Quinean observation sentence like "A rabbit" or "Rabbit." Such sentences do not contain singular terms, while sentences reporting visual inspection of objects do. The criterion to be proposed presently would be inapplicable.

gular term referring to the image.[7] The translation proposal is this: Call a representation a representer refers to or would expressly refer to if she were reporting her own achievement, a representation that is the representer's object, or briefly an *objectual representation.* Contrast this case with perception. The visual representation there involved is looked right through, and a minimally appropriate report of the episode would not have to mention it.[8] The visual representation, in this case, is not the representer's object: it is a *non-objectual representation.*

The proposal has advantages over both the transparency-opacity metaphor and the functional explication. The metaphor is deceptive. Taken literally, not metaphorically, transparency and opacity are properties that both have the characteristics of being object-dependent and being dispositional. By object-dependent I mean that an object's transparency or opacity supervenes on its structural properties and is not a matter of a perceiver's attitude towards it. Whether your windshield is still transparent or whether your window-blinds are still opaque is a matter of objective fact, not your way of looking at them. Moreover, if you call an object transparent, you don't say that it is such that you actually see through it, but rather that one can do so, and likewise for opacity. Both properties thus are dispositional. Accordingly, the philosophical discussion of transparent representations has concentrated on whether transparent representations are transparent in a weak or a strong sense, i.e. whether they can or cannot be attended to (see Kind 2003), but this problem is a mere artifact of the dispositional feature of transparency. The phenomenal description of the role of representation in perception as opposed to, say, imagery does not require this feature. In perception, the representation is not the representer's object in the sense that she simply does not attend to it. Whether such attending is possible should just be left open in the phenomenal description, but when such description uses the transparency metaphor it unduly prejudges the question. Likewise for the feature of object-dependence. If a representation's objectual or non-objectual character depended on the representation, it

[7] By a minimally appropriate report I mean one that forms a grammatically correct English sentence. Thus, someone reporting the result of inspecting a beagle image with: 'The beagle's ears are oblong' would be reporting appropriately, someone reporting it with: 'Oblong', thus suppressing a singular term, wouldn't. What is not required for an appropriate report is the cognizer's correct identification of the image as such by the singular term. She may leave it to our understanding of the task that with "the beagle" she does not refer to a dog but to a mental dog image.

[8] A similar characterization is attempted by Leeds for transparency: "... the acid test of a transparency view is what sorts of properties and objects the view takes us to be aware of in perception: equivalently, what objects and properties the words in our perceptual sentences refer to. So long as these are properties of the perceived physical objects, and so long as introspection does not produce awareness of ... other objects and properties, then the view is by my lights a transparency view." (2002, 111)

would be implausible that the representer can change that character by a mere shift of attention—which is what proponents of weak transparency advocate. But this feature, again, is an artifact of the metaphor and does not come from the phenomenon we initially sought to describe by means of it. The objectual/non-objectual distinction is free of these misleading connotations and exactly captures the phenomena—enough to justify the unsightly neologism.

The objectual/non-objectual distinction also has advantages over the functional conception with which I have set in. If something fulfills or fails to fulfill a function it must be actively used to that end. If, on the other hand, something only has a function, it is unclear whether it is in actual use—and the question of success or failure just remains open—or whether it sits there for potential use only. 'Having a function' can be read as expressing a potency just as the dispositional 'transparent' and 'opaque' do. Now, in the examples from the psychology of vision, those characterizations of a representation that I interpreted as their being transparent clearly did not express a potential use for seeing the visual object but an actual one. So, characterizing such a representation as one that *has the function* of making something a *representandum* misleads in the sense that this function might presently not be exerted while we want to express that it actually is. Exerting a function, thus, is more specific than just having one but less so than fulfilling or failing to fulfill one. The cognizer's attempt of having an object as the *representandum* could founder on the representation exerting but not fulfilling its function, but it could also do so on other reasons—because other functional entities do not fulfill their functions or because the object to be the *representandum* simply is not there. A transparent representation still could serve its purpose in the whole process of making something the cognizer's *representandum*, and for it to be used thus—as a contribution to this end—it has to actually have that certain property circumscribed as transparency. Calling it by the name of non-objectuality avoids the metaphor and the ambiguous function terminology insofar as it exactly clarifies the status of the representation relative to the cognizer when she actually uses it to function for the intended purpose. All the same holds for the objectual case, the one where something (a visual object or an image) is characterized as having the function of simply being the object of inspection. As the actual status of a mental entity vis-à-vis the cognizer in her cognitive activity is what we ultimately want to characterize, the function terminology is too unspecific.

There is an objection, both obvious and fundamental, to the proposal as a whole. What—one might wonder—*is* an objectual representation? What makes an entity a cognizer's *representation* of an object, if it is the cognizer's *object* itself? It seems that for something to be a representation it must be involved in the process of representing something so that there is

an object as potential or actual *representandum*. Not only is such an object missing in the objectual case, the representation is the cognizer's object instead. This leads one to either introduce another mediating representation representing the original one as a new *representandum*—an obviously unacceptable regress—or else to question how something can be at once a representation and a representer's object.

The objection is answered from my last remark about the function terminology. Both objectual and non-objectual representations are entities that have the function to represent something else, but it is only the non-objectual which exerts this function, i.e. is used for the purpose of representing something else. The objectual representation, having and exerting the function of serving as an object of inspection, does not exert the representational function it also has. The awkward consequence of this is that objectual representations do not actually but only potentially represent. However, the awkwardness dissolves when we realize that we know objectual representations from the realm of the non-mental very well and we call them representations in quite a different sense. Consider a map of New York City that you actually use for finding your way about. Assume that your own explicit reports clarify that you refer to lines and patches in the map (recall the suggested criterion) so that the map is objectual in the sense proposed. The map *is* a representation of the city, but it does not presently exert its function to make the city your *representandum*. Instead, it is appropriate to say that, as you cannot adopt a bird's eye view on the city, the map replaces that view, and if only you could adopt the former you would gladly forego the latter. The map is a representation in the sense of being a proxy for the real thing. For a contrast, consider a mental representation in perception again. To assume that this representation represents in the sense of being a proxy for its *representandum* is a time-honored move in the philosophy of perception—sometimes attributed to Descartes—and it is erroneous because a perceptual representation cannot occupy the place its *representandum* would occupy, if only it were available. The *representandum*, even if available, could not take the place its representation occupies; hence the latter is not a proxy of the former. And this clarifies in what sense objectual representations do and do not represent. Objectual representations are not actual representations in the sense that they do not exert the function of making some object the representer's *representandum*. They are actual representations in the sense that they exert the function of replacing a representer's *representandum*, serving as its proxy. And *vice versa* for non-objectual ones.

It can be anticipated that images and mental models alike will be classified as objectual representations below. But are images and models really representations that do not actually represent? Well, they do represent in the sense of being proxies for unavailable *representanda*, but they do not

in the sense of exerting the function of actually making entities *representanda*. Indeed, philosophical and psychological attempts to conceptually capture the roles of images and models provide more evidence for their proxy status. Mental models engendered "in the absence of a visual world" have been characterized as "surrogate representations" or as "substitutes for the represented." [9] Such models thus are viewed as proxies for *representanda* that are unavailable (perhaps because they do not exist). If they were they would be in the very place now occupied by their proxies and serve the function now taken over by the latter. This function, of course, is the one of being an object of inspection—which implies that the entity in question is the representer's object or is objectual to her. The representational function, on the other hand, is not exerted. This follows from the presupposition that the model takes the place of the potential *representandum because* the latter is unavailable. If it were it would stand in the model's stead. Thus, if we assumed that a model does both—is a placeholder for some unavailable *representandum* and at the same time actually functions to represent that *representandum*—it would have to be meaningful to further assume that the *representandum*, if available, could occupy the representation's place *and* actually represent something, to wit itself. Since this is not meaningful, the model itself does not actually represent as long as it is a proxy. An entirely parallel reflection could be repeated from Kosslyn's remark that having a mental image is seeing in the absence of a visual object, because what the remark intends to capture is that the former functions as a proxy for the latter.

2. Pictorial and symbolic representations

A second and much more obvious distinction among representations is the familiar one between symbolic and pictorial. A symbolic representation, a string of symbols, say, is related to its intended *representandum* by convention and does not need to share any of its properties. On the other hand, a pictorial representation *does* share properties with the *representandum* which is essential to its representational function and does not need to be related to the *representandum* by convention. Now, I have introduced the distinction between non-objectual and objectual representations in cases of pictorial representation, but it cuts across the other one between pictorial and symbolic representation. Thus, there can be, at least in principle, pictorial as well as symbolic representations that are either non-objectual

[9] See the final discussion in Garrod (forthcoming) for the former and Vosgerau (2006, 262) for the latter expression.

or objectual to the representer. A most important type of symbolic representation is propositional representation, and the objectual/non-objectual distinction offers an important means to describe different relations a cognizer may have to such a representation. Consider the following little theory about propositional attitudes which, though arguable, is a reasonable theory and crucially depends on that distinction. Everyday propositional attitude reports can be ambiguous. A first-person report like "I believe that this animal is a lizard" obliges me to the consequence that there is something which I take to be true, because of the semantics of 'believe.' But a third-person report like "Alicia believes that this animal is a lizard" does, despite the semantics, not commit me to the claim that there is something Alicia takes to be true. Alicia's belief is about this animal, not about a proposition, although it involves one. If she uttered her belief she would be just uttering, not mentioning a proposition. My report about Alicia commits me to assume that she would assent to "Do you believe that this animal is a lizard?" but not that she actually does. My own situation could be exactly like Alicia's, but it isn't as I am reporting my belief and for that purpose must mention the proposition believed. What happens in both cases is easily described. Alicia, if my report of her is true, is in a state that surely involves a propositional representation, but one that need not be objectual to her. My report about myself, if true, does involve a propositional representation, but one that is objectual to me—which is evident from the fact that one phrase of my report mentions it. Alicia may be in different states that involve the propositional representation in question—believing, doubting wishing, and so on—but this does not imply that she has a certain attitude toward it. The case changes, if she attends to these states, e.g., in order to report them. Now this little theory, regardless of whether it can be further defended, draws its intelligibility from the fact that distinct relations a representer can have toward a representation—it can be non-objectual to her, a mere means to another end, or objectual, an end at some point of cognitive activity and a means only for further such activities—are illustrated by propositional representations.

I have anticipated what the conceptual machinery set up so far aims at. Mental models, like perceptual representation and images, are pictorial representations (in the wide sense just explained), and they are, unlike perceptual representations, but exactly like image, objectual representations. There may be important differences between mental models and mental images (and some cognitive psychologists insist that there are crucial differences), but here is an important characteristic that images and models share: They are objectual representations or proxies for their *representanda*. The claim is evidenced quite easily. A model is produced, manipulated and inspected in the same sense the image is. Nevertheless, a mental model is produced with the conscious intention to represent something—it is, after

Mental Represenation	**Objectual**	**Non-Objectual**
Symbolic	Propositional Representation in Self-Ascribed Propositional Attitude	Propositional Representation
Pictorial	Mental Image, Mental Model	Perceptual Representation

all, a representation. Indeed, typical tasks involving mental models are not carried out for their own sake, but in order to learn something about the model's *representandum* which, for whatever reason, is unavailable. The model thus is the cognizer's object, but in a straightforward sense also is a representation. It is an objectual representation in the sense explained. Accordingly, if we chart both distinctions in a little table, models and images share one of the boxes. They are both pictorial and objectual representations.

3. Telling apart models and images

The previous cross-classification puts models and images in the same boat. Can they be told apart by pure reflection? An ongoing and inconclusive discussion casts doubt on the possibility.[10] However, a distinction is possible, if, again, we attend to the representations' functions.

Several pairs of features have been invoked for distinguishing both classes of mental representation, the most important of which is the presence or absence of abstract elements. Intuitively, the distinction can be motivated thus: Though neither image nor model participate in a process of perceptual representation (as both are objectual representations), an image is more intimately related to such a process. This intuition now is easily explicated. A representation, I argued, cannot have a function in its own generation and such generation is the representer's first step in perception as well as imagery. The difference of both processes originates in the different functions the representation is given in them. The representation generated, be it used in a perceptual or imagery process later on, does not differ in both cases. Its use in the next step just remains open, initially. Thus, a potential imagery representation is one that could as well be used—or, speaking after the fact, could have been used—as functioning in a perceptual process,

[10] Some suggest that mental models are but a special class of mental imagery (e.g., Denis1991*b*, 117; Gottschling 2006), while others insist that the two classes are distinct (e.g. Johnson-Laird 1996, 114-120).

though it is given and eventually exerts a different function in the imagery process. Though it does not actually function in a perception process it is content-wise suited for such a function. In contrast, a mental model, likewise produced in a first step and then given a function very similar to an image (i.e. made an object of inspection) in a second step, could not have been put to a perceptual use because it is content-wise unsuited for this purpose.

Plausible as the distinction sounds it is misleading because it concerns the contents of representations—which are a matter of sheer contingency. This is to say: There may be a difference here, but it is certainly not clear-cut. What the reference to a representation's content conceals is that the criterion of differentiation is, again, its function. To see this, consider the claim that models can contain abstract elements but images cannot. The easy explanation just suggested is that a representation containing an abstract element is unsuited for use as a perceptual representation just because the abstract element contained cannot represent any feature of a perceived object. Now, a potential image is suited for perceptual use, so no potential image can contain abstract elements, nor can, *a fortiori*, an actual image.

The distinction thus described plainly seems to be one of content. However, this impression vanishes if we consider how abstract elements in mental models are characterized. A typical example, taken from Johnson-Laird (1996, 117), is to decide what follows from "All the guests are ticket-holders" and "Lisa is not a ticket-holder." Subjects can draw an informative conclusion from a model only when adding new individuals who are not guests and either are or are not ticket-holders. Thus, new elements must be introduced and must be characterized by means of negation. The characterization can be realized through an arbitrary new feature or element, pictorial or symbolic, for negation, but even if an image is employed "the image itself does not do the work of negation." The element can fulfill the function of negation only if there is a clear procedure "for interpreting the image—that is, for mapping a negative sentence into the image, and for mapping the image back into a negative sentence" (Johnson-Laird 1996, 116). The description is apt. It is the function which an element of the representation is given during the use of the latter which decides over its abstractness, not that element as such. Thus, a potential model may consist of elements every bit as pictorial as the ones of a potential image. None of a representation's elements can inherently be characterized as abstract, and nothing in its content prejudges whether an element is to be used abstractly or not. The use to which representations are put decides whether they are models or images, hence they are distinguished via function, not content.

The impression that the distinction could be drawn by reference to a representation's content can be explained. In most cases, cognizers from the start produce images and models with contents suited for either image or modeling tasks. So, representations may be produced with contents suited for different purposes, but nevertheless it is our understanding of this intended use that decides for or against model or image, not the content in itself. Such use, then, is different for potential images and models. The abstract elements in a model enable it to stand for a prototypical situation from a whole class of situations with certain fixed properties. In this sense a model represents a class of situations—in contrast with an image that represents a single situation (see Johnson-Laird 1996, 120). To sum up, a representation that a representer gives the function of being an object of inspection or treats objectual is either a model or an image. If such an objectual representation is given the function of representing (in the sense of being a proxy for) a single situation, then it is an image. If it is given the function of representing a prototype from a class of situations, then it is a mental model.

4. A consequence: No mental models in perception

One immediate consequence of the previous conceptual exercise is that contentions by other authors about the role of mental models in perception are false. Johnson-Laird, in his 1983 book, writes the following philosophical aside: "Human beings, of course, do not apprehend the world directly; they possess only an internal representation of it, because perception is the construction of a model of the world" (Johnson-Laird 1983, 156). To say, as Johnson-Laird does here, that perception involves the construction of a mental model, means to say either that a mental model is used in a case of perception, but is not an objectual representation—in opposition to my proposal—or that a mental model is objectual and simultaneously is used as a mediating representation in perception—in opposition to the original proposal that representations in perception are transparent or non-objectual.

Both options are unattractive. Take, first, the idea that mental models are not objectual representations. The proposal means, we recall, that an entity is a representation (a proxy), but nevertheless the cognizer is directed towards it and can mention it. This characteristic guarantees that we can understand how the cognizer can inspect or even mentally manipulate the representation. Namely, such inspection or manipulation entails that the inspecting or manipulating cognizer is directed toward what he or she intends to inspect or manipulate. Without objectuality it becomes

unintelligible, how cognizers can do such things to representations, or, for that matter, can consciously operate on them in any way.

Take, second, the idea that mental models are involved in perception, but are, at the same time non-transparent or objectual. According to this proposal, there are no non-objectual or transparent representations in perception. However, non-objectuality initially guaranteed that representations have and exert a certain function in perception, while at the same time perception is an immediate contact with real objects in the sense that the perceiver is consciously directed only toward these objects, not toward representations of them. Without this feature perception must be reinterpreted: From being a direct contact with real objects into an indirect one via inferences from representations. But it is a mistaken description of the phenomenon of perception to say that it involves any inferences from the perceiver's immediate objects—her representations.

Nothing in Johnson-Laird's general account of the cognitive uses of mental models hinges on their presence in perception. Characteristically, the representations that theoreticians take to be involved in perception are not, in normal cases, inspected or manipulated by the perceiver and, accordingly, there is no need to interpret them as models. Admittedly, the idea that perception does involve models draws a certain attraction from neurophysiological and psychological insights into the enormous processing power involved in perception, especially in vision. David Marr, in his seminal 1982 book, describes the tremendous amount of processing of visual information at different stages of vision. Accordingly, he assumes that perception involves reference to the processed representations and inferences from them to real objects—entirely in accord with Johnson-Laird's remark. Marr writes in fact that "the true heart of visual perception is the *inference* from the structure of an image about the structure of the real world outside" (68 Marr 1982, my emphasis).

If indeed perception involved inferences from representations to objects it would be an open option to take the original representations to be models (as Johnson-Laird suggests), instead of (as Marr thinks) images. But it is implausible to claim that perception involves such inferences and, as far as I can see, nothing in Marr's theory corroborates his idea. According to him, the representations involved in perception, say, the "primal sketch" or the "2 1/2 D sketch," are not described as if the perceiver *consciously* operated on them or manipulated them. On the contrary, Marr's wording suggests that the processing of representations works on the computational level and is entirely unconscious to the perceiver (e.g., when he says that "a number of processes operate on the primal sketch to derive a representation ... of the geometry of the visible surfaces;" Marr 1982, 42). And this is, of course, the more plausible interpretation of these processes. On

the whole, there is no need to confuse objectual representations like models and images with non-objectual ones like perceptual representations.

5. An application: Mental models and propositional representations

Finally, I wish to show that there is psychological evidence for the objectual and non-objectual character of mental representations, i.e. evidence from typical mental model experiments. Mani and Johnson-Laird, in a series of experiments,[11] investigate situations where subjects tend to construct mental models from a set of descriptions as opposed to cases where they stick with just the descriptions. The authors describe experiments where subjects receive two sets of descriptions of spatial relations among simple objects (spoon, knife, plate, fork, cup). The first set determines the spatial relations among all objects, thus allows for the construction of an unambiguous mental model of the situation. The second set is ambiguous as to spatial relations and allows for more than one model. The subjects remembered the spatial arrangement much better in the case where one unambiguous model could be constructed, but they could not remember the descriptions themselves as well as in the case of several possible models. Johnson-Laird comments:

> Evidently, subjects tend to remember the gist of determinate descriptions better than that of indeterminate descriptions, but they tend to remember the verbatim detail of indeterminate descriptions better than that of determinate descriptions. This 'cross-over' effect is impossible to explain without postulating at least two sorts of mental representation. A plausible account of the pattern of results is indeed that subjects construct a mental model of the determinate descriptions, but abandon such a representation in favour of a superficial propositional one as soon as they encounter an indeterminacy in a description. (Johnson-Laird 1983, 162)

From the present view-point, Johnson-Laird's interpretation of his results can be taken one step further. Obviously, what he calls "propositional representation" must be present in both cases. After all, the subjects are given descriptions from which they must extract meanings in the form of propositional representations. So, both kinds of representation are involved in both situations. In the case where the propositional representations allow for constructing an unambiguous model, these representations tend to become transparent or non-objectual, and the subjects are directed, through them, toward the model they suggest. In the ambiguous case where no sta-

[11] See Mani & Johnson-Laird (1982), Johnson-Laird (1983, 160-62; 1996, 95-96).

ble model can be constructed, subjects have a tendency either to stick with the propositional representations or return to them. These representations then become what they are directed to, i.e. they become objectual.

The example involves two kinds of representations and the non-objectual ones here are the propositional representation. Models cannot, by definition, be non-objectual. It may be possible that a representation loses its objectuality and becomes non-objectual. However, in this case, the mental model ceases to be a model. Usually, when a cognizer arrives at a result by manipulation of a model, the result is explicitly transferred by an inference to an eventual real object the model is a proxy for. This, in fact, is the case in the communication of an explanation. It has been proposed (by Garrod forthcoming) that a model, instead of always being a "surrogate" or proxy, can function as an "interface" between the subject and a real situation. This proposal suggests that a model in this case does not remain objectual, but slips over into a non-objectual representation. What the cases described by Garrod show, however, is that such a representation loses its very status as a model and becomes a non-objectual representation, an actual means to represent a situation in the real world.

References

Denis, M. (1991*a*), *Image and Cognition*, Harvester Wheatsheaf, New York.

Denis, M. (1991*b*), Imagery and thinking, *in* C. Conoldi & M. McDaniel, eds, 'Imagery and Cognition', Springer, New York, pp. 103–131.

Garrod, S. (forthcoming), Referential processing in monologue and dialogue with and without access to real world referents, *in* E. Gibson & N. Pearlmutter, eds, 'The Processing and Acquisition of Reference', MIT Press, Cambridge, MA.

Gottschling, V. (2006), Visual imagery, mental models and reasoning, this volume, pp. 211–235.

Johnson-Laird, P. (1983), *Mental Models*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P. (1996), Images, models, and propositional representations, *in* M. de Vega, M. Intons-Peterson, P. Johnson-Laird, M. Denis & M. Marschark, eds, 'Models of visuospatial cognition', Oxford University Press, New York, pp. 90–127.

Kind, A. (2003), 'What's so transparent about transparency?', *Philosophical Studies* **115**, 225–244.

Kosslyn, S. (1994), *Image and Brain: The Resolution of the Imagery Debate*, MIT Press, Cambridge, MA.

Kosslyn, S. (1995), Mental imagery, *in* S. Kosslyn & D. Osherson, eds, 'An Invitation to Cognitive Science. Vol.2: Visual Cognition', MIT Press, Cambridge, MA, pp. 267–296.

Leeds, S. (2002), 'Perception, transparency and the language of thought', *Noûs* **36**, 104–129.

Mani, K. & Johnson-Laird, P. (1982), 'The mental representation of spatial descriptions', *Memory and Cognition* **10**, 181–187.

Marr, D. (1982), *Vision: A Computational Investigation in the Human Representation of Visual Information*, Freeman, San Francisco.

Spelke, E., Gutheil, G. & Van de Walle, G. (1995), The development of object perception, *in* S. Kosslyn & D. Osherson, eds, 'An Invitation to Cognitive Science. Vol.2: Visual Cognition', MIT Press, Cambridge, MA, pp. 297–330.

van Gulick, R. (1988), 'A functionalist plea for self-consciousness', *Philosophical Review* **97**, 149–181.

Vosgerau, G. (2006), The perceptual nature of mental models, this volume, pp. 255–275.

This Page is Intentionally Left Blank

Mental Models and the Mind
Carsten Held, Markus Knauff, Gottfried Vosgerau

The Perceptual Nature of Mental Models

Gottfried Vosgerau [1]

Department of Philosophy, Universität Tübingen [2]

Abstract
In the first comprehensive formulation of the theory of mental models, Johnson-Laird proposes several constraints on any psychological theory of reasoning. He argues that his theory fulfills these constraints due to two properties of mental models: structure preservation and naturalness. However, during the elaboration of his theory over the last decades, especially the central property of naturalness was not paid much attention to. It hence has to be questioned if the theory in its present form still possesses the explanatory power originally claimed. In this chapter, I will outline an interpretation of structure preservation and naturalness within a philosophical framework. This leads to the claim that mental models are structures partially isomorphic to what they represent and that they contain exclusively perceptual relations. I will close with some proposals for refining the theory of mental models, such that the originally proposed constraints can be met (again). Only this refined version can stand as a true alternative to theories of mental logics.

[1] I am a member of the research project "Self-consciousness and Concept Formation in Humans" lead by Prof. Albert Newen, sponsored by the VokswagenStiftung. This work was partly made possible also by the German Academic Exchange Service (DAAD), whose fellowship allowed me to visit the philosophy department of NYU. I am more than grateful for the helpful comments by Carsten Held, Vera Hoffmann, Albert Newen, Giuliano Torrengo, Laura Toulouse, Klaus Rehkämper, Stefan Wölfl, and Alexandra Zinck.
[2] E-mail: vosgerau@uni-tuebingen.de

1. The Basic Idea of Mental Model Theory

Mental models have become a widely used concept in various disciplines. Unfortunately, the use of the term varies across the different applications, such that a common notion or even a core meaning is difficult to find. In order to describe the nature of mental models, it therefore seems fruitful to re-explore the basic ideas that lead to the theory of mental models provided by Johnson-Laird (1983).

The theory of mental models was mainly developed as an alternative to theories of mental logics. All kinds of mental logics require mental representations in a specific format, namely a propositional format or "Language of Thought." According to this view, information is encoded in propositions upon which rules can be applied to process new information.

The crucial difference between mental models and mental logics is the representational format underlying reasoning. There are, above all, two properties that—according to Johnson-Laird—are responsible for the supremacy of mental models over mental logics: structure preservation and naturalness.

In the following section, Johnson-Laird's conception of structure preservation and naturalness will be outlined. The second section will provide a sketch of a philosophical framework, in which these concepts could be integrated and further explained. The third section will combine both views and lead to some refinement of the concept of mental models.

1.1. Small-scaled models of external reality

The idea that mental representations are "small-scaled models of external reality" can be traced back at least to Craik (1943). The basic intuition is that mental representations are structured, and that this structure mirrors the one of the *representandum* (the represented object or situation). Therefore, the effects of changes affecting the model can be directly interpreted as effects that would occur in the real situation if the according changes had been performed. This gives the mind the power to simulate possible actions or processes without carrying them out. The whole process of modeling—including changes and the interpretation of the effects—leads to new information about the represented, which is called reasoning. If, for example, I have a physical model of the constellation of the sun, the earth, and the moon, I can understand the phenomenon of solar eclipse without having seen it (in real size). Mental models are understood very much like such small-scaled models we often use in explaining physical phenomena.

The crucial difference to propositional formats of representation as proposed by theories of mental logic is that no logical rules have to be learned.

The reasoning process can be explained without referring to any presupposed logic. Quite on the contrary, taking into account certain capacity limitations, the "failure" of human reasoning in certain situations can be explained while at the same time the principle logic competence of human reasoners and the development of abstract systems like logic and mathematics becomes conceivable (cf. Johnson-Laird 1983, 125, 144f). For the sentences "The apple is to the right of the banana" and "The banana is to the right of the cherry" the following model can be constructed:

cherry banana apple

It can now be directly "read" from the model that the apple is to the right of the cherry. For this inference, the logical rule of transitivity is not needed as it would be the case for propositional representations (cf. Johnson-Laird 1995, 1000). Moreover, the competence of humans to reason according to such rules is explained.

Mental models are hence complex representations that share their structure with their *representandum*. The explanatory power of mental model theory relies—according to Johnson-Laird—on the fact that mental models are structure-preserving representations. If they lacked this property, the competence of logical reasoning would depend on abstract and sophisticated notations. These notations would have to be learned in some mysterious (or at least implausible) way. Moreover, structure preservation ensures sound reasoning. Logical thinking emerges in mental models. Therefore, there is no possibility of applying logical rules falsely and hence no possibility of having a correct mental model but failing to reason correctly (leaving capacity limitations aside).

I will now turn to a second feature of mental models that is necessary for the explanatory power Johnson-Laird believes them to provide: naturalness.

1.2. Naturalness

As indicated above, one major advantage of mental model theory—as seen by Johnson-Laird—is that no logical rules have to be learned in order to reason logically sound. This advantage would vanish if mental models contained abstract features themselves. To evade this problem, Johnson-Laird therefore describes mental models as being natural. This means that they do not involve "sophisticated mathematical notations" (Johnson-Laird 1983, 93). Euler Circles, which represent sets as circles in a plane, for example, are hence bad candidates for mental models. Alternatively, a set is (usually) represented by some characteristic members in mental models (cf. Johnson-Laird 1995). Therefore, "a *natural* mental model of discourse has a structure that corresponds directly to the structure of the state of affairs"

(Johnson-Laird 1983, 125). The constraint of structure preservation hence does not suffice to provide *natural* representations: There has to be a *direct* correspondence. Unfortunately, Johnson-Laird is rather obscure about how to spell out directness.

If a mental model directly corresponds to the modeled situation, the relations in the model have to correspond directly to the modeled relations as well. This leads to the even more central requirement that the relations between elements of a mental model have to be natural as well. In the above example, it is quite clear that the relation 'to the right of' is represented not by an arbitrary symbol or another "abstract notation" but—in the natural way—by itself. What exactly "natural" relations are, as opposed to abstract relations, is not stated by Johnson-Laird himself.

On the contrary, Johnson-Laird introduces several abstract notations. Indeed, the notations he applies vary across his writings. For example, he introduces a symbol for negation, which is clearly a "highly sophisticated notation." In order to be structure-preserving and natural, however, a mental model should contain representations exactly for the elements that are part of the represented situation. Everything that is *not* part of the represented situation is simply omitted in the model as well. Hence, there is no need for negations to be expressed in mental models.[3] I will discuss some of Johnson-Laird's more recent remarks on this issue in section 3.2.

The criterion of naturalness is closely linked to the explanation of learning to reason. Theories that presuppose logical rules or notions have to explain how these rules or notions can be learned. "If a theory proposes that a sophisticated logical notation is used as a mental representation, then it should offer some account of how such an apparatus is acquired." (Johnson-Laird 1983, 66) It seems implausible that these notations are innate since most people have significant difficulties in learning logical and mathematical systems. Since mental models are natural, they do not contain sophisticated logical notations. In this way "[t]he theory solves the central paradox of how children learn to reason" because it shows that "[i]t is possible to reason validly without logic" (Johnson-Laird 1983, 145). Hence, the notion of naturalness carries the burden of providing an unproblematic basis for this learning ability.

The only reasonable cognitive ability that can be presupposed before inferences are learned is perception. Therefore, I conclude that the only way to understand the notion 'natural' properly is to read it as 'grounded in perception.' Hence, the relations contained in mental models have to be found in perception as well. Examples for such relations are surely 'to the

[3] In my view, practically all such abstract notions introduced by Johnson-Laird can be eliminated or viewed as abbreviations. However, a discussion of his notation would lead too far into details that are hardly of concern to the basic ideas.

right of,' 'brighter than,' 'sweeter than,' but also kinesthetic relations like 'being moved by me' or 'being moved by some external force.' This does not mean, however, that mental models are themselves perceptual in the sense that they could be objects of perception. Nor does it follow that mental models are modal-specific. Take for example spatial mental models: They can contain only spatial relations which are perceptual. Since there is no percept with spatial relations alone and there are many modalities in which spatial relations are perceived, purely spatial mental models are neither perceptual nor modal-specific. Moreover, as we know from neuroscientific research, perception can achieve a very high level of abstraction.[4] Indeed, the transition from a pictorial stage to an abstract language-like stage often proposed in developmental psychology (following Piaget) does not necessarily involve the construction of new abstract representations: When a child learns to apply the (already abstract) representations containing only perceptual relations to represent other than perceptual problems (e.g. using spatial relations to represent temporal problems; see also Johnson-Laird 2001), then she will exhibit "abstract skills" without employing new formats of representation. Whether these abstract perceptual relations are suitable for the description of the reasoning power of trained adults, or whether some mechanism for abstracting even further must be introduced, will not be discussed here. My aim is merely to describe the basic idea of mental models. For this purpose, I have focused on two constraints—structure preservation and naturalness—which are crucial for the explanatory power of mental model theory.

2. The philosophical account

2.1. The property of being a representation

In philosophy, the discussion about the nature of representations has a long tradition, going back at least to Plato and Aristotle. The attempt of this section is not to give a summary of this discussion, but rather to present a quite loose framework for the discussion of the special case of mental models.

The two major problems every representation theory has to face are the explanation of the asymmetry of the representation relation and the explanation of misrepresentation. If R is a representation of X, it usually follows that X is not a representation of R. The architect's model is a representation of the house, whereas the house is not seen as a representation of

[4] For example in vision, as described by Marr (1982).

the model. Moreover, there are misrepresentations, that is, cases in which a representation fails to work properly. If the fuel gauge is broken, the needle will misrepresent the amount of fuel in the tank. A theory of representation which explains only ideal cases while not taking into account failures would be highly inappropriate.

Causal theories hold that a representation is caused by its represented. In these theories it is very difficult to give an explanation of misrepresentation, for there is no such thing like miscausation. Especially for mental representation the so-called disjunction problem arises: If a horse erroneously causes a cow-representation, then this cow-representation cannot have the content 'cow' because it is not caused by a cow. Rather, if cow-representations can be caused by horses, then they have the content 'cow or horse.' In the end, this leads to the conclusion that (almost) every mental representation has a disjunctive meaning. If this were the case, our interaction with the world would be rather poor for we could not distinguish between horses and cows. Even refined versions like the one of Fodor (1987, 1994) do not seem to evade this problem. Fodor proposes a nomic relation to hold between the representation and the *representandum*: The horse-caused cow-representation is asymmetrically dependent on cow-caused cow-representations; if there were no cow-caused cow-representations, neither there would be horse-caused ones, but not the other way round. However, in order to have cow representations, a cognitive system has to have the ability to discover an eventual mistake, i.e. it has to be able to tell cows from horses (Fodor outlines this requirement for the case of frogs, which he takes to have black-moving-dot-representations rather than fly-representations; see Fodor 1994, 196f). To be able to distinguish cows and horses means to have different mental representations of cows and horses. Hence, according to Fodor, a cognitive system can have horse-caused cow-representations only if it is able to have cow-representations. Therefore, the nomic relation that is necessary for a representation to have a certain content can only be established if the system already has representations with this certain content. There seems to be no easy way out of this circle and so Fodor's solution of the disjunction problem fails.

Theories of similarity are ruled out because of two reasons: Firstly, most similarity relations[5] are symmetrical, and therefore fail to account for the asymmetry of the representation relation. Secondly, even if there are non-symmetrical similarity relations,[6] there will be much more objects being similar to each other without representing each other. Nevertheless, there is a long tradition of similarity theories for mental representation (e.g. Aristo-

[5] There are a lot of similarity theories which differ in the definition of similarity (cf. Cummins 1989).

[6] See, for example, Demant (1993).

tle, Hume, the early Wittgenstein), involving different similarity relations. In fact, structure preservation—one of the two basic features of mental models—is a special kind of similarity, often called isomorphism. One of the clearest philosophical articulations of isomorphism theories has been given by Cummins (1996). Introducing the example of a robot that is able to navigate through a maze, he argues that the robot's representation has to be isomorphic to the actual maze: Whatever the representation looks like in detail, it has to guide the movements of the robot; if the movements and hence the representation are not isomorphic to the maze, the robot will not succeed. I will argue in a similar vein in section 2.2. However, Cummins concludes that isomorphism is sufficient for all mental representations, which is certainly too strong in two respects: Firstly, not all mental representations have to function in that way, and second, isomorphism cannot be sufficient for the representation relation to hold since a) it is symmetric, and b) not everything isomorphic to something else represents it. Nevertheless, I will come back to isomorphism in the next section.[7]

A third type of theories is built by the so-called functional theories. They hold that a representation becomes a representation by taking over the functional role of the represented (for example Dretske 1994, Millikan 1994, Cummins 1989). Following our intuition, a representation is something that stands for something else. Standing for something else is not an inherent property of objects. A tennis ball, for example, does not stand for anything by being a tennis ball. However, it can stand for the moon (while the earth is represented by a soccer ball, for example) in a certain context. It is then *used* as a representation for the moon. In the context of showing the constellation of earth and moon, the tennis ball becomes a representation of the moon because it takes over the role that the moon plays in the "real" constellation. A representation is hence an entity that is used to stand for something else in a certain context.[8] It becomes a representation for this or that by taking over the role of this or that.

In more detail, for mental representations this means that behavior is normally described as some sort of function mapping some inputs to outputs. Especially in reasoning, the output is (the utterance of) a belief (new information not given in the premises). The reasoning process in our example (see page 257) can be described as a function mapping the premises

[7] Goodman (1976) famously argued against similarity theories of representation, instead proposing a conventional account. However, his discussion focuses on works of art, whereas my focus is on (special kinds of) mental representation. Since convention presupposes several users which can (implicitly) agree on some convention, this account is not suitable for mental representations (they have only one "user") and will not be discussed here.

[8] For this reason, representations are always tokens. Speaking of representations as types must be viewed as an abbreviation if not as mistaken.

Table 1
Mental representations as substitutes

let a, b, c be the apple, the banana, the cherry, resp.
let α, β, γ be the mental representation of the apple, the banana, the cherry, resp.
let R be the relation between the fruits
let P be the relation between the representations of the fruits
there is a function f: $R(a,b)$, $R(b,c) \mapsto$ belief that $R(a,c)$
the substitution $\left[\frac{P(\alpha,\beta)}{R(a,b)}\right]$, $\left[\frac{P(\beta,\gamma)}{R(b,c)}\right]$ yields:
f: $P(\alpha,\beta)$, $P(\beta,\gamma) \mapsto$ belief that $R(a,c)$

to a conclusion. However, in the world (about which we reason), there is an according function doing much the same. If I set up the situation of the apple, the banana, and the cherry, I will also come to believe the conclusion by seeing it. Hence, mental representations of situations can be described as stand-ins (substitutes) for the real situations in a specific function. For this reason, they are representations of these situations. In the above example, the function maps two situations in the world (the apple lying to the right of the banana and the banana lying to the right of the cherry) onto the conclusion "The apple is on the right of the cherry" (see Table 1). Mental representations can take the place of the real situations in this function. When they are substituted by mental representations, the functional roles of the situations are taken over by the mental representations, allowing the reasoner to come to the same conclusion without looking at the world. Representations can hence be characterized as substitutes for the real situation in a specific function. [9]

The account sketched so far is a quite plausible and appealing one, for it straightforwardly explains the asymmetry of representation. However, it does not offer a satisfying explanation of misrepresentation. [10] In the next paragraph, I will try to show that this is due to the fact that a crucial feature of representations is completely overlooked by functionalists.

Although a representation becomes a representation only by being a substitute for the represented, it is obvious that there are better and worse

[9] The behavior described as a function must not be confused with the function of the represented object, for example the nourishing function of the fly for the frog. Of course, these functions cannot be taken over by mental representations. Nor am I talking about the function of the representation, i.e. to stand for the represented; talking about representation in this way only states the problem instead of giving an explanatory account (*pace* Millikan 1994).

[10] Millikan (1986), for example, explains misrepresentation with abnormal circumstances. However, it remains an open question exactly what normal circumstances are.

representations for one and the same thing. A schematic railway map of Germany is certainly a representation I can use to travel Germany by car. Nevertheless, a much better representation for this purpose is a road map. The reason for this is, intuitively speaking, that the road map contains more relevant information than the railway map. It does so independently of the user. Hence, there are some features of the road map which make it a suitable candidate for using it as a representation of the roads of Germany. Functionalistic approaches to representation overlook the fact that there is an important relation between the representation and the represented object. However, this relation is not enough to establish a representation relation. Nevertheless, it determines an object's suitability for being used as a representation for a certain entity. There may be simple representations which do not stand in any (relevant) relation to the represented. However, most representations we apply are complex representations: models, sentences, pictures, etc. A representation fails, i.e. is a misrepresentation, if it is used as a representation in spite of being inadequate. A map of France will be a misrepresentation if I use it to find my way through Germany.

Since a representation is a substitute for the represented, it takes over its functional role. However, the output of the function will not be accurate if the representation is not adequate. In other words, it must be able to take over the functional role; otherwise, the output of the function will not be reliable. Therefore, there must be a relation between the representation and the represented object that is independent of the functional roles. I will call this relation the adequacy relation. It is likely that there are different adequacy relations, as there are different kinds of representation. In the case of linguistic symbols, for example, the adequacy relation seems to be convention, whereas convention is a rather implausible candidate for the adequacy relation of mental representations.

I have analyzed the representation relation as consisting of two parts: the taking over of the functional role, and the adequacy relation, which holds between the representation and the represented. There seem to be different kinds of representation that differ exactly in respect to the adequacy relation (models, sentences, pictures, ...). In the following, I will confine myself to the discussion of the adequacy relation between a *model* and its *representandum*.

2.2. The relation between a model and its represented

Following Craik (1943) and Johnson-Laird (1983), a model preserves the structure of its represented. It is able to react to changes in the way the *representandum* would when undergoing the according changes. A prerequisite for this ability is that the model contains parts that represent parts of the

modeled situation. These parts have to be connected in the same way as their "real" counterparts. This approach to structure preservation remains quite intuitive.

In the philosophy of science, scientific theories are often viewed as models. Although there is a debate on whether models can be characterized as being isomorphic to reality, many authors defend this view.[11] In psychology, there is a long tradition of discussing whether mental representations can be viewed as isomorphic to their *representanda* or not. However, there have been quite a few attempts to define the different concepts properly (cf. Palmer 1978, Gurr 1998). Therefore, I will start with the mathematical notion of isomorphism.

In mathematics, structures are sets over which one or more functions and/or relations are defined. Two structures $\mathfrak{A}$ and $\mathfrak{B}$ are said to be isomorphic if there is a bijective mapping I between the $a_i \in \mathfrak{A}$ and the $b_i \in \mathfrak{B}$, such that
- for each function f: $I\left\langle f^{\mathfrak{A}}(a_1, \ldots, a_n)\right\rangle = f^{\mathfrak{B}}(I\langle a_1\rangle, \ldots, I\langle a_n\rangle)$ and
- for every relation R: $I\left\langle R^{\mathfrak{A}}(a_1, \ldots, a_n)\right\rangle$ iff $R^{\mathfrak{B}}(I\langle a_1\rangle, \ldots, I\langle a_n\rangle)$.[12]

The definition requires that for each member of one set there is exactly one corresponding member in the other set. Moreover, for every function defined on one set there must be a function defined on the other set that picks out the corresponding element given the corresponding arguments, and for every relation that holds in one set, there must be a relation holding for the corresponding elements of the other set. Now, one of the two structures can be a certain part of the world, for example a house. In the architect's model of the house (which is then the other structure), every piece of the house can be assigned a corresponding piece of the model, and every relation between those elements of the house will correspond to some relation in the model. However, since there are more elements and more relations in the world than in the model, this example does not satisfy the definition: Not every single brick is modeled. I will return to this matter shortly. Nevertheless, taking isomorphism as a requirement for models, it follows that if X is a suitable model of Y, then for every element of Y there must be exactly one element of X corresponding to it. Johnson-Laird expresses this requirement by the idea that mental models represent each individual taking part in a situation by a single part of the model. The appropriate model for the sentence "The apple is on the left of the banana" hence involves two tokens, one for the apple and one for the banana (see Figure 1).

However, the mathematical notion of isomorphism is too strong a requirement for most models. It is obvious, that, for example, the architect's

[11] For a discussion see French (2002).
[12] Cf. Ebbinghaus et al. (1992, 49).

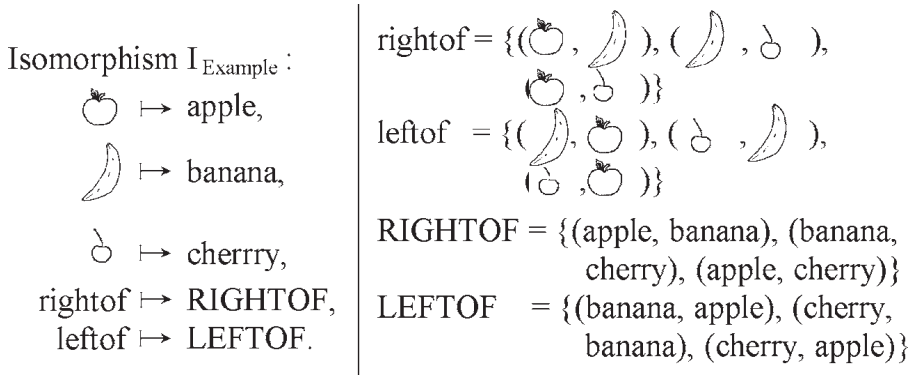


Fig. 1. The isomorphism between the world and the model in the example (see page 257)

model of a house does not have as many elements as the real house. Similarly, there are many relations between the apple and the banana in the real situation (concerning their color or size, for example) which are very unlikely to be contained in a mental model used in reasoning about the spatial relations. It is thus useful to introduce the notion 'relevant part of a structure,' which is determined by the usage of the representation. If I want to reason about the spatial relation of fruits, a mental model containing spatial relations will suffice. On the other hand, if I want to decide which fruit to eat, there certainly will be more relevant relations to represent (for example, is sweeter than). More technically, the relevant part of a structure is determined by the function in which the *representandum* is involved (see page 261) based on what relations and functions are taken as arguments. [13] If $\mathfrak{A} = \langle A, R^{\mathfrak{A}}, f^{\mathfrak{A}} \rangle$ is the structure of a situtation, the relevant part of the structure $\mathfrak{A}'$ will consist of the same set A, a subset of the relations $R^{\mathfrak{A}}$ and a subset of the functions $f^{\mathfrak{A}}$. These subsets are the sets of relations and function which are taken as arguments by the function in which the model plays its role. We can therefore speak of a partial isomorphism which holds between the relevant part of the represented structure and the full

[13] I take it for granted that models represent situations and not just the world; furthermore, I take a situation to contain certain objects. Hence, a mental model indeed has to represent every object of the situation, but not every object in the world.

model. [14]

According to this definition, models are structures that are isomorphic to the relevant part of the structure of the represented object; the relevant part is determined by the function in which the represented object plays its role. Although often proposed, the weaker criterion of homomorphism cannot do the job for two reasons. Firstly, homomorphism does not involve a bijective mapping. Therefore, although the mapping of the *representandum*'s parts to parts of the model may be unequivocal, the inverse mapping may be not. Hence, the output of the function would not be guaranteed to be applicable to the represented object. Secondly, since homomorphism does not specify parts of structures, even very small parts can establish homomorphism. Therefore, for each structure there are many trivially homomorphic structures that are much too unspecific to be called models. The second point applies as well to partial isomorphism (as introduced by Bueno 1997, French & Ladyman 1999), unless the parts are otherwise specified (as I have done above). Moreover, partial isomorphism in the sense presented above allows for an evaluation of models: A perfect model is a model which is isomorphic to the whole relevant structure of the represented. If it is isomorphic to more or to less of the relevant structure (or even contains parts that are not shared by the represented), it is a bad model. A model containing too much will be more difficult to construct and to manipulate; therefore, it will make reasoning less effective. On the other hand, a model containing too little obviously will not be able to take over the functional role adequately, since relevant pieces of information are missing. Moreover, the 'more' and 'less' of the deviation from the perfect model can be measured: If the relevant part of the represented structure $\mathfrak{A}'$ contains $m_{\mathfrak{A}}$ relations and $n_{\mathfrak{A}}$ functions, and the model $\mathfrak{B}$ contains $m^{+}_{\mathfrak{B}}$ relations and $n^{+}_{\mathfrak{B}}$ functions fulfilling the conditions of the partial isomorphism, and $m^{-}_{\mathfrak{B}}$ relations and $n^{-}_{\mathfrak{B}}$ functions not fulfilling the conditions, then the deviation δ^{+} of the model from $\mathfrak{A}'$ can be defined as $\delta^{+} = \left|(m_{\mathfrak{A}} + n_{\mathfrak{A}}) - (m^{+}_{\mathfrak{B}} + n^{+}_{\mathfrak{B}})\right|$, and the amount of irrelevant information δ^{-} as $\delta^{-} = m^{-}_{\mathfrak{B}} + n^{-}_{\mathfrak{B}}$. The adequacy ϵ of the model can then be defined as

[14] I tacitly assumed that a mental model is not just a physical entity that happens to be within someone's cranium. Rather, there are certain operations executed on the model. These operations do not operate on all physical relations and properties of the realizer of the model. Only those relations and properties that are relevant for the operations are taken to be relations and properties of the mental model (see also Palmer 1978, for a discussion of operations and mental representations). Although it is true that there are many relations in the description of mental models that are not representing (e.g. 'banana' has more letters than 'apple'), a mental model contains only representing relations. Therefore, the partial isomorphism involves the whole structure of the model and not just a relevant part of it.

$$\epsilon = \left(1 - \frac{\delta^+}{(m_{\mathfrak{A}} + n_{\mathfrak{A}})}\right)\left(1 - \frac{\delta^-}{(m_{\mathfrak{A}} + n_{\mathfrak{A}})}\right).$$

This leads at least to a relative measurement of model adequacy, i.e. it allows for an evaluation of models. [15]

Isomorphism is a relation between structures. Hence, a model is itself a structure, i.e. a set over which functions and relations are defined. Thus, the appropriate model of the example (see page 257) can be written as

$$\langle \{a, b\},\ \text{leftof} = \{(a, b)\},\ \text{rightof} = \{(b, a)\} \rangle .$$

The crucial point is that a model does not represent the relations involved as symbols (or labels); it itself contains relations which hold between its elements regardless of whether it is used as a model or not. Since the relations have the same logical features [16] as the relations of the real situation (see the definition of isomorphism), they exhibit the same structure. This is why the isomorphism theory is so attractive: It explains straightforwardly why our conclusions are correct (given that we have a good model and no capacity limitations). Nevertheless, as argued for in section 2.1, isomorphism theories have to be embedded in a functional theory in order to explain the phenomenon of mental representation; partial isomorphism is just one part of the representation relation for models, namely their adequacy relation.

One possible objection to isomorphism addresses the representation of non-existing situations: In reasoning, I usually construct models of situations that are merely supposed to have but do not actually have any counterpart in the world. To what should these models be isomorphic? To answer this question, let me recall that isomorphism is a relation between structures. The mental model is hence not isomorphic to a situation but to the structure of a situation. Structures themselves are abstract entities (consider, for example, the structure of natural numbers with the relation '$\geq$'). The structure of a non-actual situation is as unproblematic a notion as the set of natural numbers is. Therefore, it is possible to have an adequate model of the situation described by the sentence "There is a golden mountain in Africa," since there is a straightforward notion of a structure of this situation, even though it is not an actual situation. To illustrate this, it might be helpful to note that we can agree on structural "facts" about non-existing entities (e.g., we can agree that unicorns have four legs).

[15] This measurement may not reflect the cognitive effectiveness of mental models, since it assumes that irrelevant information is as hampering as missing relevant information, which is of course an open empirical question. This question could be addressed by introducing a weight to the amount of irrelevant information.

[16] With 'logical features' I refer to features of relations such as transitivity, symmetricity, reflexivity, etc. The definition of isomorphism implies that correspondig relations also have the same logical features.

Thus, the representation of non-existing situations is explained in my picture without committing myself to some problematic ontology (like realism about possible worlds, for example).

Stenning (2002) points out that mental models are not special in respect to isomorphism. Equally, other forms of deduction systems such as Euler Circles and fragments of natural deduction systems stand in this relation to their represented objects. They are all "members of a family of abstract *individual identification algorithms*" (Stenning & Yule 1997, 109). Therefore, structure preservation is not the crucial feature of the theory of mental models that distinguishes it from other theories of reasoning; rather, the constraint of naturalness plays the distinctive role. However, I will not go deeper into this debate but rather discuss some major implications of my analysis, particularly the use of symbols in mental models.

3. The structure of mental models

3.1. The explanatory power of mental models

Considering the implications of the isomorphism condition and the condition of naturalness, we can conclude that mental models are structures, which are isomorphic to the relevant part of the structure of the represented, and which contain only relations that are based (i.e. also found) in perception. In particular, this means that for every object taking part in the represented situation there is one token in the mental model. These tokens stand in different perceptual relations to each other. Every relation in the model has an according counterpart in the situation that has the same logical features. Hence, if a mental model is perfect in the sense that it is isomorphic to the relevant structure of the represented, then sound logical reasoning "emerges" from this representational format. Failures occur due to the use of bad models and due to capacity limitations of working memory (cf. Johnson-Laird & Byrne 1991, Johnson-Laird 2001). Therefore, reasoning with mental models does not presuppose the knowledge of logical rules. On the contrary, it explains why people are able to reason logically and develop such formal systems as logic and mathematics. Moreover, the riddle of how children acquire reasoning skills is solved insofar as the only mechanisms presupposed are perception and memory.

The requirement of natural relations together with the requirement of isomorphism is crucial for the explanatory power of mental models. Isomorphism ensures soundness and natural relations ensure learnability. In the terms of Palmer, mental models are "intrinsic representations," i.e. they are "naturally isomorphic" to the represented (Palmer 1978, 296f). However,

Palmer calls this kind of isomorphism natural because the logical "structure is preserved by the nature of corresponding relations themselves" (Palmer 1978, 297).[17] In contrast to this, 'natural' has a more specific meaning in my interpretation of Johnson-Laird (1983): The relations are not only natural in Palmer's sense but also natural as opposed to artificial or abstract, which means perceptual. Only under this interpretation, the theory can be said to throw light on the problem of learnability of logical reasoning.

Since other theories of reasoning propose mental representations that exhibit partial isomorphism (cf. Stenning 2002), the constraint of structure preservation is not special to the theory of mental models. The various "individual identification algorithms" (Stenning & Yule 1997, 109) turn out to be equivalent, i.e. there is no algorithm belonging to this family that can compute more than another. Moreover, it is not clear what kind of processes determine the difficulty of a specific reasoning task. In mental model theory it is the number of models that have to be constructed. On the other hand, for fragments of natural deduction systems, Stenning (2002) points out that it is not clear that the number of rules to be applied is a sensible measure of task difficulty. Therefore, mental model theory cannot be tested against other theories of reasoning belonging to the same family unless there are crucial features other than partial isomorphism. Johnson-Laird (1983) claims that his theory explains more than just sound inferences and fallacies of reasoning: The problem of learnability is solved by constraining mental models to be natural mental representations. Hence, the specific explanatory power of this theory, which distinguishes it from others, relies on the naturalness constraint. As I argued (in section 1.2), this constraint cannot explain how children are able to learn logically sound reasoning unless it is interpreted as 'grounded in perception.' Therefore, if the constraint of naturalness is given up or weakened, the specific explanatory power of mental models is lost and the theory becomes eventually indistinguishable from other theories of reasoning.

Nevertheless, Johnson-Laird changed his view on the naturalness con-

[17] The idea behind non-intrinsic representations is that the logical properties of relations can rely on other sources than the intrinsic relations between elements of representations. For example, the sign '$\geq$' can be defined to be a transitive relation; however, the sign itself is not intrinsically transitive. Nevertheless, since it follows from the mathematical definition that every isomorphism is intrinsic, the distinction is rather one of the source of the logical features of the relations involved. I already pointed to the assumption that the relations in a model are partly defined by the operations executed on them (see footnote 14). Hence, if the mental model is taken to be the structure that is (partly) defined by the operations (and not just the physical realization), then mental models become trivially intrinsic representations. However, the perceptual relations I talk of are equally (partly) defined by the operations executed on them (not every physical property of some neuronal signal has to be relevant for its processing). Therefore, the distinction between intrinsic and non-intrinsic isomorphisms does not affect my argument.

straint, now claiming mental models to contain symbols for abstract notions. In the last subsection I will discuss some of his recent remarks about the nature of mental models in more detail. I will sketch an alternative view compatible with the research done so far and with the explanation of learnability.

3.2. Symbols in mental models

Discussing the "existential graphs" of Ch. S. Peirce, Johnson-Laird (2002) draws some implications for his theory of mental models. I will pick out his fourth implication "[...] that you cannot have an iconic representation of negation." He concludes: "Hence, no visual image can capture the content of a negative assertion" (Johnson-Laird 2002, 84). 'Iconic' is used here in the sense of Peirce, i.e. a sign is an icon of something if it (visually) resembles the designated entity. Since there is nothing resembling negation, there cannot be iconic representations of negation. This point can easily be extended to perceptual relations (and properties), since negation is not perceivable. Therefore, negation can only be designated with the help of a symbol, i.e. a sign that bears its meaning due to convention. Accordingly, "mental models therefore use a symbol to designate negation" (Johnson-Laird 2002, 85). There are several difficulties with that view: convention in mental representation, learnability, and the scope of negation.

The first problem is a general problem of symbols (in the Peircian sense) as mental representations. Symbols are signs that gain their meaning by convention. Their meaning is fixed by some agreement of the sign users (which is often established by usage). However, mental representations cannot be conventional since there is only one single user. This single user cannot make any agreement and hence cannot fix the meaning of any symbol.[18] If functionalism is true, then every mental representation gains its specific "meaning" (what it stands for) by having a specific causal role within the system. This causal role cannot rely on an agreement by others. Therefore, mental representations can never be symbolic in the Peircian sense.

Negation is a sophisticated logical notion, and hence every theory of reasoning that introduces the notion of negation "should offer some account of how such an apparatus is acquired" (Johnson-Laird 1983, 66). Johnson-Laird does not offer such an account and therefore does not meet his criteria for theories of reasoning. It might be true that the notion of negation has to be learned at some point in order to develop the full adult reasoning skills. However, if so, we need an explanation of when and how

[18] This argument is closely related to the famous private language argument of Wittgenstein (1922).

it is learned. Otherwise, the distinctive feature of naturalness in mental model theory vanishes.

The third problem arises when we look at what mental representations represent. They represent situations (state of affairs), real ones as well as supposed ones. For this reason, a mental model contains elements corresponding to elements of the modeled situation and relations (and properties) corresponding to relations (and properties) in the modeled situation. Everything that is a part of the situation will be represented by something in the mental model. Everything that is not found in the situation will have no counterpart in the mental model. So far, there is no need for representing negation because there is no negation "in the world," and mental models are partially isomorphic to the "world."

Negation is a truth-functional operator of sentences, i.e. only sentences can be negated.[19] It states that the so-called proposition, which is expressed by the sentence, is false, i.e. that the situation described by the sentence is non-actual. Since mental models stand for situations, it is not clear why there is any need to represent negation *within* a model. Rather, the whole model should be negated, i.e. there should be a possibility to make clear that the situation represented by the model is non-actual. There are different relations in which a subject can stand to representations of situations: She can believe that p, wish that p, fear that p, and so on (where 'p' can be substituted by some English sentence). These different relations are called propositional attitudes. Propositional attitudes are often explained as functional roles: The belief that p can be explained as the mental representation of p that plays a certain functional role for the thinker's behavior. If I search for my pencil on the desk, for example, I will do so partly because I believe it is there. The belief that my pencil is on the desk hence plays a certain functional role in my behavior and can therefore be characterized as a belief. Likewise, believing that something is true, probable, possible, false, or supposed can be characterized as different propositional attitudes. The difference between a mental model that represents some real situation and a mental model representing only a supposed situation is therefore a difference in functional roles. A supposed situation will not change my behavior in the way an actual situation does. In the same way, negation (of whole models) can be explained in terms of functional roles. Therefore, no representation of negation is needed in mental model theory. Of course, the acquisition of the ability to differentiate between different functional roles has to be explained. However, this need for explanation is not restricted to reasoning theories.

Let us take a look at other sentence operators. If there is—as stated by

[19] Adjective phrases are usually analyzed as abbreviations for sentences ("the nice house" for "the house is nice"). Therefore, adjectives can be negated as well.

Johnson-Laird—a need for a negation symbol, why is there no need for conjunction, disjunction, and implication symbols? A conjunction is represented simply by putting the two required models into one. Everything that stands in one mental model is conjunctively connected (Johnson-Laird 2002, 87). A disjunctive sentence, on the other hand, is simply resolved by representing each of the possibilities in a separate model (Johnson-Laird 2002, 86). Implications are treated in the same way.[20] There is no need for symbolic representations of these operators because they relate different models and not different elements of models. The same holds for negation: Because negation operates on mental models there is no need for a symbol within mental models. Of course, there is still need of some form of "mental negation." However, it is explained with the help of specific functional roles of the model. In the same way as a believer does not have to have a symbol for belief in order to have beliefs,[21] a reasoner does not have to have a symbolic representation of negation in order to reason with negated models.

Taken together, introducing symbols for negation into mental models contradicts both the constraint of structure preservation and the constraint of naturalness. Moreover, it is not obvious why this has to be done. Quite on the contrary, there are straightforward ways of introducing negation into the theory without a need to presuppose representations of negation. Therefore, if the theory of mental models should be a real alternative to other theories of reasoning, the use of symbols in mental models has to be abandoned. Otherwise, its distinctive explanatory power is lost, since introducing symbols is not compatible with the naturalness of mental models.

4. Conclusion

In the first comprehensive formulation, the theory of mental models (Johnson-Laird 1983) is introduced with two basic constraints on mental models: structure preservation and naturalness. Both constraints contribute substantially to the distinctive explanatory power of the theory.

Within a functionalistic frame, these basic constraints can be spelled out more precisely. A mental model stands in a certain relation to the represented situation. In order for the model to work, this relation has to be a

[20] This is possible because each implication $p \rightarrow q$ can be written as a disjunction $\neg p \vee q$.
[21] Beliefs simply affect her behavior in a certain way and are thereby characterized as such; some philosophers use the metaphor of a belief-box to illustrate this view: A representation is a belief if it is in the belief-box (as opposed to the desire-box, for example). The representation itself does not contain a symbol or any other information about its being a belief.

partial isomorphism, which assures soundness of thinking. The constraint of naturalness is not that clear in the writings of Johnson-Laird. He believes that mental model theory can solve the problem of learnability of logics. He states that the naturalness of mental models does account for learnability. Mental models are natural because they do not contain abstract mathematical or logical notions. However, if the learnability problem is taken seriously, the constraint must be even stronger. The only ability we are certain children acquire before acquiring reasoning skills is perception. Hence, the relations contained in a mental model have to be found in perception as well. Still, mental models do not have to be perceptual themselves, nor are they modal-specific.

It has been shown by Stenning (2002) that partial isomorphism is not only limited to mental models. It follows that the constraint of structure preservation is not unique to mental models. Hence, the distinctive explanatory power of mental model theory has been proven not to stem from this constraint. Therefore, the constraint of naturalness has to take over the burden of giving the theory its distinctiveness. Nevertheless, this constraint seems to play a marginal role in the later works of Johnson-Laird. He introduced many abstract notions into mental models which are clearly not perceptual. In this way, the problem of learnability is not solved by mental model theory, as it stands today, and a great deal of the theory's explanatory power is given away. Taken together, it is no longer clear what the fundamental difference is between mental model theory and other theories of reasoning (like mental logics; see Stenning 2002). Only if the constraint of naturalness is reactivated and consistently built into the theory, the distinctive explanatory power of mental model theory can be established.

Johnson-Laird was the first to stress the importance of structure preservation of mental representations. He also showed that so-called analogous representations need not to be modal-specific (like mental images) but can be quite abstract while remaining grounded in perception (see for example Knauff & Johnson-Laird 2002). However, to clearly distinguish mental model theory from other theories of reasoning in the future, the naturalness constraint must be clearly defined in psychological terms and consistently applied to the explanation of the phenomena. I have given an analysis of negation and proposed a way of omitting a negation symbol in mental models. The other abstract notions that are currently used in the theory have to be analyzed in a similar manner. Moreover, the notion of perceptual relations has to be defined in psychological (and neurological) terms; so far, this has been done mostly for visual relations. I think that this project is promising since the resulting version of mental model theory would have a very strong explanatory power that could hardly be gained by any other theory of reasoning.

References

Bueno, O. (1997), 'Empirical adequacy: A partial structures approach', *Studies in History and Philosophy of Science* **28**, 585–610.

Craik, K. (1943), *The Nature of Explanation*, Cambridge University Press, Cambridge.

Cummins, R. (1989), *Meaning and Mental Representation*, The MIT Press, Cambridge, MA, London.

Cummins, R. (1996), *Representations, Targets, and Attitudes*, The MIT Press, Cambridge, MA, London.

Demant, B. (1993), *Fuzzy-Theorie oder die Faszination des Vagen*, Vieweg, Brauschweig, Wiesbaden.

Dretske, F. (1994), Misinterpretation, *in* S. Stich, ed., 'Mental Representation: A Reader', Blackwell, Cambridge, MA, Oxford, pp. 157–173.

Ebbinghaus, H.-D., Flum, J. & Thomas, W. (1992), *Einführung in die mathematische Logik*, BI-Wissenschaftsverlag, Mannheim, Leipzig, Wien, Zürich [english translation: Mathematical Logic. New York: Springer, 1994].

Fodor, J. (1987), *Psychosemantics*, The MIT Press, Cambridge, MA, London.

Fodor, J. (1994), A theory of content, II: The theory, *in* S. Stich, ed., 'Mental Representation: A Reader', Blackwell, Cambridge, MA, Oxford, pp. 180–222.

French, S. (2002), 'A model-theoretic account to representation', *Proceedings of the PSA* (Supplement).

French, S. & Ladyman, J. (1999), 'Reinflating the semantic approach', *International Studies in the Philosophy of Science* **13**, 99–117.

Goodman, N. (1976), *Languages of Art*, Hackett Publishing Company, inc., Indianapolis.

Gurr, C. A. (1998), On the isomorphism, or lack of it, of representations, *in* K. Marriott & B. Meyer, eds, 'Visual Language Theory', Springer, New York, Berlin, Heidelberg.

Johnson-Laird, P. N. (1983), *Mental Models*, Harvard University Press, Cambridge, MA.

Johnson-Laird, P. N. (1995), Mental models, deductive reasoning, and the brain, *in* M. S. Gazzaniga, ed., 'The Cognitive Neurosciences', MIT Press, Cambridge, MA, pp. 999–1008.

Johnson-Laird, P. N. (2001), 'Mental models and deduction', *Trends in Cognitive Sciences* **5**(10), 434–442.

Johnson-Laird, P. N. (2002), 'Peirce, logic diagrams, and the elementary operations of reasoning', *Thinking and Reasoning* **8**(1), 69–95.

Johnson-Laird, P. N. & Byrne, R. (1991), *Deduction*, Lawrence Erlbaum Associates, Hove (UK).

Knauff, M. & Johnson-Laird, P. N. (2002), 'Visual imagery can impede reasoning', *Memory and Cognition* **30**(3), 363–371.

Marr, D. (1982), *Vision: A Computational Investigation in the Human Representation of Visual Information*, Freeman, San Francisco.

Millikan, R. G. (1986), 'Thoughts without laws; cognitive science with content', *The Philosophical Review* **95**, 47–80.

Millikan, R. G. (1994), Biosemantics, *in* S. Stich, ed., 'Mental Representation: A Reader', Blackwell, Cambridge, MA, Oxford, pp. 243–258.

Palmer, S. (1978), Fundamental aspects of cognitive representation, *in* E. Rosch & B. L. Lloyd, eds, 'Cognition and Categorization', Erlbaum, Hillsdale, NJ, pp. 259–302.

Stenning, K. (2002), *Seeing Reason*, Oxford University Press, Oxford.

Stenning, K. & Yule, P. (1997), 'Image and language in human reasoning: A syllogistic illustration', *Cognitive Psychology* **34**, 109–159.

Wittgenstein, L. (1922), *Tractatus Logico-Philosophicus*, Routledge & Kegan Paul, London.

This Page is Intentionally Left Blank

Index

This Page is Intentionally Left Blank