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: $I\left\langle f^{\mathfrak{A}}\left(a_1,\ldots,a_n\right)\right\rangle = f^{\mathfrak{B}}\left(I\left\langle a_1\right\rangle,\ldots,I\left\langle a_n\right\rangle\right)$ and
– for every relation R: $I\left\langle R^{\mathfrak{A}}\left(a_1,\ldots,a_n\right)\right\rangle$ iff $R^{\mathfrak{B}}\left(I\left\langle a_1\right\rangle,\ldots,I\left\langle a_n\right\rangle\right)$.[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).

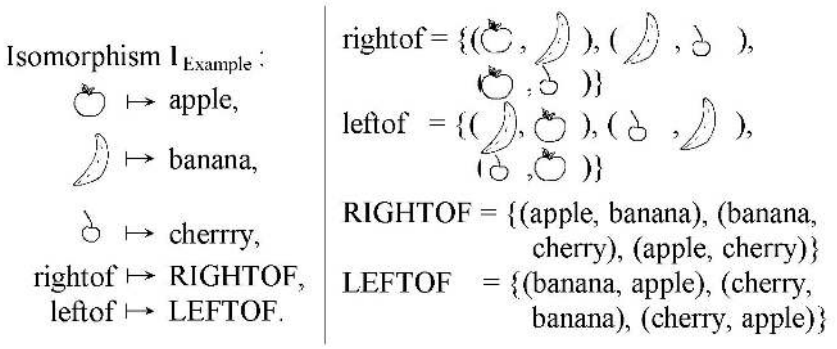

Fig. 1. The isomorphism between the world and the model in the example (see page 257)

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 $\delta^{+} = \left|(m_{\mathfrak{A}} + n_{\mathfrak{A}}) - \left(m^{+}_{\mathfrak{B}} + n^{+}_{\mathfrak{B}}\right)\right|$, 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.