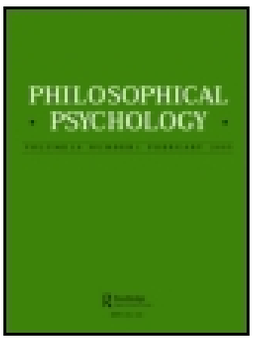


Philosophical Psychology

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: http://www.tandfonline.com/loi/cphp20

Inferential basing and mental models

Luis Rosa

To cite this article: Luis Rosa (2016): Inferential basing and mental models, Philosophical Psychology, DOI: 10.1080/09515089.2016.1266318

To link to this article: http://dx.doi.org/10.1080/09515089.2016.1266318

PHILOSOPHICAL PSYCHOLOGY, 2016
http://dx.doi.org/10.1080/09515089.2016.1266318

Routledge
Taylor & Francis Group

Inferential basing and mental models

Luis Rosa

Munich Center for Mathematical Philosophy, Ludwig-Maximilian University, Munich, Germany

ABSTRACT

In this paper, I flesh out an account of the inferential basing relation using a theory about how humans reason: the mental models theory. I critically assess some of the notions that are used by that theory to account for inferential phenomena. To the extent that the mental models theory is well confirmed, that account of basing would be motivated on empirical grounds. This work illustrates how epistemologists could offer explications of the basing relation which are more detailed and less empirically risky.

ARTICLE HISTORY
Received 4 May 2016
Accepted 13 November 2016

KEYWORDS
Basing relation; inference; mental models; psychology of reasoning

1.

What is it for one's belief to be *based on* one's reasons? And what is it for one's belief to be based on one's reasons *in the right way* (in such a way as to earn positive epistemic status)? Contemporary epistemologists have been dealing with these questions for quite some time now, and common trends on the topic can be found in the ever-growing literature. Here, however, epistemologists inevitably touch on empirical matters.

Take for example what is perhaps the majority position on the nature of the basing relation: that it is a *causal* relation. When it comes to being based *in the right way*, the proposal is that there is something like *non-deviant* causation of beliefs, where the distinction between deviant and non-deviant causation is grounded on normative considerations.[1] To the extent that the defender of the causal theory also assumes that we *do* have justified beliefs and knowledge, she is also committed to the empirical hypothesis that our beliefs result from certain types of causal chains (but not others). Or consider the proposal that a second-order belief to the effect that one's reasons are *good reasons* to believe that φ is a necessary condition for one's belief that φ to be based on those reasons in the right way. Again, to the extent that epistemologists who defend this theory also assume that we do mostly have justified beliefs and knowledge, they are also committed to the empirical hypothesis that humans usually hold beliefs about what is a good reason for what, *ergo* that justification- or knowledge-conferring reasons are transparent to the believer's mind. (The epistemologist may well be free from the aforementioned empirical commitments, of course, if she is to doubt or to suspend judgment about whether we do hold justified beliefs, or about whether we do have knowledge).

Not only are basing-theorists mostly committed to empirical hypotheses, their accounts of basing are very general accounts: details about how exactly the basing relation is supposed to take place in the subject's cognition are left out of the picture. For example, when it comes to the causal account, is the causation between reasons and belief supposed to be enabled by hard-wired neural pathways (and if not, then how)? Is long-term memory necessarily involved in the causal process? When it comes to

CONTACT Luis Rosa luis.rosa@lrz.uni-muenchen.de

the second-order belief account, must the relevant second-order attitude be held by the subject *before* she forms the doxastic attitude that is supposed to be grounded on reasons? Or is the second-order belief involved as in a process of rationalization? The degree of generality is not a problem per se – but we may find out that the actual processes of belief-formation that take place in our cognition do not quite meet the epistemic standards we have expected them to from the armchair. So knowing the specifics should be of real interest to the epistemologist.

We are not hereby implying that the epistemologist herself should flesh these details out and try to find out whether they accurately describe our cognitive processes (that is presumably a job for the cognitive psychologist/neuroscientist). But as long as our theories about the basing relation are not at all in an empirical vacuum, we should certainly look for the best theories in psychology and neuroscience and try to integrate them into our epistemological theories. There is nothing wrong with risking the possibility of empirical disconfirmation (as we mostly are in this area of investigation) – but why not back our theories up with the relevant data and the best explanatory hypotheses? Doing so should also help us to come up with more detailed accounts of the basing relation.

Quite obviously, however, we cannot expect that by just "adding some science" to our epistemological theories we will be readily backing up our empirical commitments with solid scientific evidence and giving rise to finer grained accounts of the relevant notions. For example, without some important conceptual clarifications, we are probably not going to get rid of the ambiguity with which the psychologist/neuroscientist sometimes uses certain mentalistic and epistemic terms. In general, a critical assessment of the scientific theory is mandatory.

In this paper, we illustrate how we could fill in the details about the basing relation in an empirically informed way. More specifically, we show how a particular psychological theory about human reasoning could give rise to a more detailed explication of a certain type of basing relation.

2.

For the sake of tractability, then, let us begin by restricting the scope of our investigation to the relevant type of basing relation: it will be about *inferential* basing. Furthermore, the type of inference that will interest us here is *deductive inference*. Deduction is the clearest case of transmission of knowledge or justification through inference. To engage in the task of fleshing out empirically informed accounts of basing, a certain level of specificity is required. There are many different types of cognitive phenomena that go under the label of *basing*, for example, forming beliefs on the basis of perception and forming beliefs as a result of reasoning (consider an empirical account of how perception works and of how perceptual input makes it into our doxastic systems on the one hand and an account of how humans reason on the other). Of course, we will find similarities between these different cases of basing. But one of the points of using hypotheses from cognitive psychology to empirically inform epistemology is that the armchair notion of basing is *too general,* in the sense that details about how basing takes place in each type of situation are left out of the picture.

So the cases that will interest us here are cases in which the subject's belief is based on certain reasons *R* via deductive inference. There are at least two ways in which this type of basing may figure in our doxastic lives. First, it may take place when the belief that is said to be based on reasons *R* is *formed* through deduction from *R*. This is *belief-generation* inferential basing. Second, the inferential basing relation between reasons *R* and the target belief may take place when the subject already holds the belief, but she deploys *R* in support of it (maybe in addition to other reasons she had to hold it). This is *belief-support* inferential basing, and it occurs *after* the target belief has been formed by the subject. Belief-support inferential basing may occur either as a result of *forward reasoning* (the subject draws inferences from her reasons and notices that φ, which is already believed by her to be true, is included among those inferences) or as a result of *backward reasoning* (the subject inquires into what premises could give support to φ, which is already believed by her to be true, and ends up finding acceptable premises that give support to φ).

In both cases, belief-generation and belief-support, we can still maintain that the subject's target belief is inferentially based on reasons *R*, even though time has passed since the initial basing relation took place (either the one responsible for forming the belief in the first place or the one responsible for making the belief reason-enhanced). Of course, in order for that to be the case, some additional conditions must be met. More obviously, in order for it to be true that the subject's belief is still based on reasons *R*, both the belief and the reasons *R* must still be held by the subject. Perhaps some counterfactuals must be true here as well, such as counterfactuals establishing a dependence relation between the target belief and the reasons on which it is based or counterfactuals about what the subject would do if asked why the target proposition is true (the one proposition that constitutes the content of her inferential belief).[2]

So what area of investigation in the natural sciences could be relevant to the project of fleshing out an empirically informed account of the deductive inferential basing relation? Certainly one place to look is the cognitive psychology of reasoning (see Manktelow, 1999).

Here we should draw a distinction between: (i) theories about *what* reasoners are trying to accomplish when they reason (or about what is being computed when a cognitive system is reasoning), and (ii) theories about *how* reasoners reason (or about how things are computed when a cognitive system is reasoning). For example, "rational analysis" theories of reasoning, most prominently Bayesian approaches (see Oaksford & Chater, 2001), are theories of the former type. They purport to explain a certain class of cognitive phenomena by making reference to the reasoner's *goals* (e.g., the goal of gaining substantial information), and then they show how this is nicely modeled by a probabilistic framework. Rips (1994) "psychology of proof" theory, on the other hand, is a theory about how we reason. It says that when we are faced with certain deductive tasks (e.g., determining whether a certain conclusion follows from certain premises) we build *mental proofs* in our working memories. The proof steps would be licensed by inference rules that are in some sense available to the cognitive system: the system would scan the premises in its working memory and would look up rules that can be applied to those premises (based on grammatical structure), thus outputting new conclusions back into working memory. Theories of this type are usually called "mental logic" theories (where "logic" is really just supposed to mean a syntactically governed machinery).

Given our present goal (i.e., to flesh out a more detailed proposal about *how* beliefs can be deductively based on reasons and to do so in an empirically informed way), we will focus on a particular theory about *how* we reason, not on theories about *what* we are trying to accomplish when we reason.

3.

The theory we are going to make use of here is the *Mental Models* (MM) theory of reasoning, which was initially developed by Johnson-Laird (1975).[3] The basic idea is that we reason by constructing mental models of situations and reading off conclusions from those mental models (in experimental reasoning tasks, the relevant situations are described through explicitly given premises or vignettes). Reasoning depends on our ability to construct situations or possibilities through cognitive resources as widely used as the power to represent things using spatial relations, abstraction, memory and language comprehension (the nature of mental models is a topic we will address below). The MM theory is said to make accurate predictions about both, levels of difficulty in inference and systematic errors that people make in reasoning tasks.

When it comes to levels of difficulty, predictions are based on the claim that some pieces of reasoning require us to consider more mental models than others do: the former ones put more load on the reasoner's working memory. For example, the theory accurately predicts that reasoning with *exclusive* disjunctions of the form *p or q but not both*, is easier (faster and more successful) than reasoning with *inclusive* disjunctions of the form *p or q and possibly both*. While reasoners only need to initially consider *two* models that satisfy the former type of claim in order to reason validly (one model in which *p* is true and *q* is false and another one in which *q* is true and *p* is false), they need to consider

three models that satisfy the latter type of claim (the two previously mentioned models plus a model in which both *p* and *q* are true).[4]

When it comes to systematic errors, the theory makes predictions on the basis of what is and what is not explicitly represented in our mental models of situations. For example, consider conditionals of the form *If p then q*. According to Johnson-Laird and Khemlani (2013), when subjects are asked to reason about conditionals of this form, they initially rely on mental models that can be schematically represented thus:

p *q*

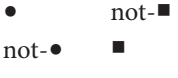
…

The first row stands for a model of the salient case in which both *p* and *q* are represented as true, and the ellipsis is a placeholder for other models in which the antecedent is false – models that are not explicitly represented by the reasoner, even though she may be aware that they are possible.[5]

So suppose a subject is given a premise of the form *If p then q* and another premise of the form *not-q*, and we ask her what follows from these premises. If the subject is to add the second premise, she has to build a mental model that is incompatible with what was explicitly represented in the initial mental model for the conditional. For example, if the consequent was *It is raining*, and the subject had initially built a mental model of a rain event for the conditional *If it is summer then it is raining*, now the subject has to consider a situation in which there is no rain and, therefore, a situation that does not overlap with the one she had considered initially. If that is right, then we should expect the subject to answer that *nothing at all* follows from the premises, for no mental model was found that satisfies both premises. And this is indeed a common response.[6] The asymmetry between modus ponens and modus tollens inferences is now a widely known phenomenon, and it is predicted by the MM theory in the way we just illustrated (see Evans, 2013).

An important principle about mental modeling is what Johnson-Laird calls the "principle of truth." Roughly put, this principle says that mental models represent only what is taken to be true – unless there is also an explicit assertion to the effect that something is not true (see Johnson-Laird & Khemlani, 2013). The principle of truth is a principle of *economy*: it reduces the amount of representation needed for the subject to reason. For example, consider the possible mental models for the exclusive disjunction *Either x is a circle or a square*:

In the first model, the only thing that is represented as being the case is that *x* is a circle, while in the second one the only thing that is represented as being the case is that *x* is a square. Yet, a fully explicit listing of the models that satisfy that exclusive disjunction would be:



The fact that there is something like the principle of truth governing our construction of mental models helps explain several experimental results, such as failure to recognize the equivalence between exclusive disjunctions and biconditionals (see Johnson-Laird & Khemlani, 2013, p. 12). In this example, reasoners will mostly fail to recognize the equivalence between the exclusive disjunction *Either x is a circle or a square* and the biconditional *x is a circle if and only if x is not a square*.

As far as a theory about *how* we reason goes, the MM theory seems to enjoy a reasonable degree of fit with the available data – not only with data from experiments involving reasoning tasks, but also with facts as widely known as (see Johnson-Laird, 2008): (1) people untrained in logic are able to perform logically sound inferences, (2) graphics and diagrams may help us reason better, and (3) purely syntactical ways of representing information makes reasoning in general computationally intractable. Another reason why we are focusing on the MM theory is that we are looking for a *unified* theory about how we reason – one that purports to explain in a similar way not only how deductive reasoning is performed, but also how inductive/probabilistic reasoning is performed. The MM theory says that we build models of situations in our heads and that we look for new instantiations of properties and relations in those models (new in the sense that they were not explicitly represented by the subject before); depending on whether the new conclusion (if any) holds in *all* or maybe in *most* of the relevant mental models, we draw conclusions that would be necessary or probable conditional on the premises.[7] So we could also use the MM theory to account for the basing relation in cases of inductive/probabilistic inferences (even though for the sake of tractability I will not try to do that here).

In order to illustrate how an empirically informed account of the inferential basing relation could be fleshed out using the MM theory, however, we need not assume that this theory explains *all* reasoning phenomena. We can assume that people reason using mental models of situations in at least a relevant class of cases, without yet claiming that the explanatory scope of the MM theory covers all the territory. Maybe sometimes we draw conclusions using a variety of heuristics or fast and frugal processes that have little to do with building mental models and reading off conclusions from them, or perhaps in some cases we literally apply inferential rules to the premises, guided purely by the syntactic properties of the relevant sentences. If really the whole class of cognitive phenomena that go under the label of "reasoning" includes not only one, but more than one type of cognitive process, then it should not be a problem that there will be more than one empirically informed account of the inferential basing relation (in fact, that is just what we should expect).

4.

The mental models theory says that we draw inferences by building mental models of situations (where in some cases the relevant situations are explicitly described through sentences), and noticing that certain things hold in those models that we did not explicitly "see" before (novel conclusions). This means that we extract information from mental models that *was already there* – so there is a sense in which the relevant conclusions are not really new (at least in the case of deduction). This cries out for important conceptual clarifications, and we better make no epistemological use of the MM theory without critically assessing the notions it makes use of. In particular, we need to get clearer on what a mental model is, how mental models differ (as *representational* items) from syntactical items, and how they relate to each other; we also need to get clearer on the distinction between explicit and non-explicit representational contents. Let us begin, then, with the concept of a mental model.

Mental models are supposed to be factual representations: they represent things as being a certain way. Perhaps we could say that they are more basic or primitive ways of representing things as being a certain way (as compared to *sentences*). But what else can we say about the nature of mental models?

Mental models are also *abstract* representations, in the sense that they abstract away from irrelevant details and represent things in common among many particular possibilities. For example, they may be rotatable 3D representations of objects that capture only their silhouettes and spatial relationships. A mental model of a plate, a fork on its left side, and a knife on its right side may represent things in common between many possibilities in which objects of those types stand in those types of spatial relations – it does not matter what colors those objects are, the material they are made of, what sort of table they are standing on, how far they are from each other, and so on. As long as the fork is represented as being to the left of the plate and the knife is represented as being to the right of it, the rest is irrelevant. Furthermore, when we build a mental model of a situation, for example as described by the sentence *The fork is on the left side of the plate and the knife is on the right side of it*, the conclusions

we may draw from it are not restricted to a single point-of-view of the situation.[8] For example, if we learn in addition that *There is a note hidden below the plate*, we may still infer that *The fork is on the left side of the note*, even though the note did not appear in the initial model.

Johnson-Laird (2008, Chapter 2) emphasizes that mental models are not images, for images are rich in detail and are presented from a fixed point of view or angle, whereas mental models need not have these properties. In the example we just mentioned, a single image of a fork, a plate, and a knife (from left to right) would represent these objects also as having certain colors (e.g., silver and white), or as being more or less aligned with each other, and they would do so only from a single point of view (e.g., as seen from above). So, for example, how could a single image be a mental model for both sentences, *The fork is on the left of the plate and the knife is on the right of it* and *There is a note hidden below the plate*? It would appear that not all mental models are images.

But even though mental models are not always *single* images (in virtue of point-of-view restrictions), perhaps they can always be taken to be *sequences* of images. And even though a single image of a situation has irrelevant details, what we might call an "image-scheme" or "sketch" will also abstract away from irrelevant details and represent what is common to a relevant set of possibilities. So, perhaps mental models are *sequences of image-schemas* that admit varying degrees of details. This is coherent with Johnson-Laird's claim that images *derive* from models (2008, p. 29).

That looks like a plausible proposal, but only insofar as we are talking about mental models of situations that can be *perceptually* represented. For example, what is a mental model for the claim that $x^2 = xx$, or for the claim that *European holidays are not getting any cheaper*, that *All hominoids are descendants of a common ancestor*, that *Some people are happy and some are grumpy*, that *First-order logic is undecidable*, or that *Metaphysicians worry about causation*? The problem is even more acute if we take mental models to be *iconic*, meaning that their structure mirrors the structure of what they represent. What kind of mental model has the same structure as the fact that first-order logic is undecidable?

In the interest of avoiding this problem, the mental model theorist has at least three positions available: (a) to hold that mental models are iconic, but that the mental models theory of reasoning is concerned only with reasoning about facts that can also be perceptually represented; (b) to hold that mental models are iconic, but that there are also iconic representations of facts that cannot be perceptually represented; (c) to hold that only a certain subclass of mental models are iconic, not all of them. Option (a) would imply a very drastic restriction on the explanatory scope of the MM theory: most of the things we reason about cannot themselves be perceptually represented (as one might conclude after reading the newspaper).

There are two ways in which one could pursue option (b). According to the first way – call it the *Platonist approach* – there are some iconic mental models whose parts represent *things* that cannot be presented to the senses. For example, in the case of mental models that satisfy mathematical statements, there would be an iconic mental model with the same structure as the fact that 2 + 2 = 4 – and *there is* such a fact in a "third realm" composed of *abstracta*, perhaps sets. According to the second way – call it the *empiricist approach* – it is still the case that every single mental model is iconic with respect to facts that can be presented to the senses, but mental models may also satisfy claims about facts that cannot themselves be presented to the senses.

Option (c) seems to be the one that Johnson-Laird opts for. As he rightly notes, concepts such as *negation*, *possibility*, and many others do not seem to have iconic counterparts, and so he allows mental models to actually contain *symbols* that relate to those concepts (2008, p. 33) (although he emphasizes that mental models are iconic "as far as they can be," 2008, p. 37). This option would imply that sometimes mental models are constituted by syntactical elements as well.

But we think this is a mistake. To be sure, one can have a mental model *of* a symbol – but that does not mean the symbol itself is part of the mental model *as a means of representing things* (in this case the symbol is the thing that is represent*ed*, not the thing that represents). Consider for example the case involving negation. There may be an iconic mental model for the claim *The square is to the left of the circle*, thus represented:

■ ●

But we cannot build an iconic model for *The square is not to the left of the circle*. So Johnson-Laird resorts to:

¬[■ ●]

But how are we supposed to interpret this? Is that supposed to mean that the subject tokens "not" at the same time at which she entertains the model for *The square is to the left of the circle*? If so, then it starts to look as if pieces of reasoning involving negation are really syntactic operations, in which case the MM theory turns into a mix of mental models and mental logics approaches to reasoning (to the extent that *that* is the interpretation it gives to models for sentences involving negation).

Now, there really seems to be no reason to deny that part of our ratiocinative activities are guided by purely syntactic operations. As we see it, however, these are not the types of cognitive performances that the MM theory is best suited to account for. It is rather best suited to account for reasoning performances in which: (i) we build iconic (non-symbolic) models that satisfy certain premises, and (ii) we realize that other sentences are also satisfied by those models, and we thus draw novel conclusions. Notice that if we allow symbols to creep into mental models, then there is no more reason to suppose that we need to consider more mental models for some premises than we need for others – the MM theory loses the explanatory power it would otherwise have if it were restricted only to reasoning that involves purely iconic models (why not use symbol-manipulation all the way in our explanation then?).

As a matter of fact, it seems that we can explain a large class of reasoning performances in such a way as it is not supposed that mental models are also made out of symbols. Consider again the negation case. What is the iconic model for *The square is not to the left of the circle*? The answer is: there are *many* such models, and we can ignore their differences – at least *as long as they are thought of as models for that sentence*. As an example, consider an iconic model thus represented:

● ■

This model satisfies the aforementioned negative sentence, and it makes absolutely no use of symbols. An equally good model would represent the square and the circle in the same vertical alignment. The principle that says that we can ignore the differences between mental models at least as long as they are thought of as models of particular sentences is a principle that the MM theory has to assume anyway. To see this, notice that there is a huge variety of ways one can build a model that satisfies the non-negative sentence *The square is to the left of the circle* as well. If you think of the center of the circle as the (0, 0) coordinate in a Cartesian plane, any model on which the center of the square is a coordinate (n, m) where $n < 0$ is a model that satisfies that sentence.

We could tell a similar (although a bit different) story about *possibility*. What could be a mental model for a sentence of the form *It is possible that p*? Again, there would be many such models: any mental model that allows for an extension in which p is satisfied, and any model in which p itself is satisfied.

That being said, as far as we are taking mental models to be purely iconic (i.e., not involving symbols) we are left with option (b): mental models are always iconic, but they can also represent facts that cannot themselves be perceptually represented. As we saw, we have two options here: a Platonist approach and an empiricist approach. The main problem with the Platonist interpretation of (b) is that it has heavy ontological commitments: suddenly the mental model theorist is committed to the existence of the third realm. Of course, we do not here purport to argue against realism about abstract objects and so on – it is just that it may bring more costs than benefits to an empirical theory about how we reason.

So perhaps we should resort to the empiricist approach to option (b). This approach says that each mental model is iconic with respect to facts that can be presented to the senses – but mental models may also represent facts that cannot themselves be presented to the senses (without being iconic *with respect to them*, but always being iconic with respect to something else). For example, any mental model that contains a representation of two objects and two other objects will also be a model that represents

four objects, thus verifying the sentence 2 + 2 = 4. The fact that 2 + 2 = 4 cannot be presented to the senses – but the fact that there are two cows here and two cows there (thus making four cows) can.[9] Or consider the sentence *Ana is happy*. We saw that we can think of mental models as sequences of image-schemas. Just as the members of the sequence may stand for different *spatial* points of view of the same object, they can also stand for different *temporal* points of view of the same object. So the fact that Ana is happy may well be represented by a series of image-schemas in which Ana behaves happily. Let us assume that these and other details are filled out by the empiricist interpretation of mental models, at least for a certain pairs of (i) facts that are represented by mental models and (ii) sentences that are satisfied by those models.

But taking this stand on the nature of mental models does not yet give us an account of the distinction between explicit and non-explicit representation, or even an account of the way in which sentences relate to mental models in reasoning. We turn to that now.

5.

We can represent a subject's mental model of a situation as follows:

$$\begin{array}{lcl} \textit{RED} & = & \{a\} \\ \textit{GREEN} & = & \{b, c\} \\ \textit{BALL} & = & \{a, b\} \\ \textit{CUBE} & = & \{c\} \end{array}$$

In the toy mental model represented here, call it *T*, *a* is an icon of a red ball, *b* an icon of a green ball, and *c* an icon of a green cube. So we use set-theoretic relations to represent the ways in which a mental model categorizes objects. As *a* belongs to *RED* in our scheme, we say that *a* matches the template *RED* in the subject's cognitive system (we use the same name for the set and the template but, purportedly, the template is in some sense stored in the subject's cognitive system, while the set is not). Let "a_T" designate the object that *a* is purportedly a picture of in *T*. So *T* represents a_T as being red and it also represents a_T as being round, etc.

If the subject only represents something to be the case through a mental model of the situation without yet representing that purported fact by means of a mentally tokened sentence, let us say that her representation is a *primitive representation*. For example, if the subject only represents a_T as being red through her mental model of the situation without yet representing that purported fact by means of a sentence like *Bally is red* (*Bally* is used as a proper name of a_T here), her representation that a_T is red is a primitive one.

Some things are *explicitly* represented in the subject's mental model of the situation, while some are only *non-explicitly* represented. For example, the fact that c_T is green is explicitly represented in our subject's model *T*, but the fact that c_T is neither red nor a ball is only non-explicitly represented. In general, the purported facts that are explicitly represented in a mental model are those whose representations *do not* depend on further factual representations in the model; and the facts that are only non-explicitly represented in a mental model are those whose representations *do* depend on other factual representations in the model. For example, *T* only *non-explicitly* represents the fact that c_T is neither red nor a ball, for its representation of this purported fact depends on (a) *T* representing the fact that c_T is not red and (b) *T* representing the fact that c_T is not a ball; furthermore, (a) also depends on (a') *T* representing the fact that c_T is green, and (b) depends on (b') *T* representing the fact that c_T is a cube. (a') and (b'), on the other hand, do not depend on further factual representations in *T* – we might call these iconic representations *atomic representations of T*. Now we can say that what is *explicitly* represented in a mental model is that which is represented in that model through an atomic representation; what is *non-explicitly* represented in a mental model is everything that is represented in it but not through an atomic representation.

Now, just as the subject may represent things as being a certain way through mental models, she may also do so through mentally tokened sentences. While *a* is an icon that stands for a_T, a rigid

designator may also be used by the subject – for example, *Bally* – in order to make reference to a_T; and while the template *RED* is somehow used by the subject in order to classify or to recognize things, she may also use a concept or a predicate, for example, *is red*, in order to attribute redness to objects. So the subject could represent the fact that a_T is red through the sentence *Bally is red* as well. If the subject represents something to be the case by tokening a sentence, then let us say that the subject has a *symbolic representation* of the purported fact thereby represented. Here, we can also draw an explicit/non-explicit distinction.

The central notion for drawing that distinction is the semantic notion of *satisfaction*: a sentence is said to be *satisfied by* a model (or a model "makes" a sentence true). For example, given that $a \in$ *BALL* in *T*, we can say that *Bally is a ball* is satisfied by *T*. This notion is an indispensable tool for the mental model theorist. We can then say that a sentence is either explicitly or rather non-explicitly satisfied (if at all) by a given mental model. A sentence is explicitly satisfied by a mental model when it is in some sense *isomorphic* to an explicit representation in that model. Let us again assume that the purported fact that c_T is green is explicitly represented in a subject's mental model *T*, where her iconic representation of c_T, or *c*, fits the template *GREEN* for her. Suppose further that our subject uses the rigid designator *Cubix* in order to make reference to c_T, and also that she uses the predicates *is green* and *is a cube* as we normally do. Then we can say that the sentence *Cubix is green* is *explicitly satisfied* by her mental model *T* – for that sentence is isomorphic to an explicit representation in *T*. But since the fact that c_T is green and c_T is a cube is only non-explicitly represented in *T* (for the representation of that complex fact depends on further representations in *T*), the sentence *Cubix is a green cube* is only non-explicitly satisfied by the subject's mental model *T*.

In what sense is a sentence *isomorphic* to an explicit representation in a mental model, so that it is explicitly satisfied by it? Since the relevant relation of isomorphism is a relation between representational items, a possible proposal is that it is a relation of *intensional* isomorphism (see Carnap, 1947, p. 59). The intension of a term is traditionally conceived as a function that maps from scenarios or situations to the extensions of that term in each scenario or situation, so that two terms have the same intension when they are mapped onto the same extensions in all possible scenarios or situations. Now consider a tuple containing the intensions of the terms that occur in a sentence *s*, where the order of the elements in the tuple mirrors the grammatical structure of *s*. Call that tuple the *structured intension* of *s* (see Lewis, 1970). Given that much, we say that two sentences *s* and *s'* are intensionally isomorphic when they have the same structured intensions. We can then try to generalize that notion: not only pairs of *sentences* can be in the relation of intensional isomorphism, but so can pairs of sentences and iconic representations. For example, the iconic representation of the fact that c_T is green in *T* would be intensionally isomorphic to the sentence *Cubix is green*, as long as icon *c* and the term *Cubix* pick out the same objects throughout different scenarios, and as long as the subject's template *GREEN* and the predicate *is green* do so as well.

Of course, structured intensions are originally supposed to mirror the *grammatical* structure of a sentence or set of sentences, and presumably iconic representations lack grammatical structure. But our generalization of the relation of intensional isomorphism need not imply that iconic representations have grammatical structure. In order for a sentence *x is F* to be intensionally isomorphic to an iconic representation *u*, *x* and *u* must have the same intensions and there must be a template *T* under which *u* falls that has the same intension as *F*, but we need not assume that <*u*, *T*> captures a grammatical structure just as <*x*, *F*> does.

We saw that, according to the MM theory, we reason by building mental models of situations and "seeing" that certain things hold in those models that we did not explicitly "see" before, thus drawing novel conclusions. In some sense, the information we extract from our mental models was there all along (at least in the case of deduction). The point of our distinctions is to help explain how this is so. So when we construct mental models of situations, there are some things that are explicitly represented and some that are only non-explicitly represented as being the case in those mental models – where this is determined by the *dependence* relation mentioned above. The novel things we "see" in our mental models are the things that were only non-explicitly represented in our mental models – but

they were represented nevertheless. Presumably, it is only through a symbolic representation that we can actually become aware that something that was only non-explicitly represented in a mental model is also the case.

6.

Let us now go back to the inferential basing relation. Our rough characterization of the MM theory would prompt the following explication of belief-generation deductive inferential basing (represented by the relational predicate "is deductively based$_g$ on"), for cases involving beliefs with *symbolic* contents:

> **(BG)** S's belief Bp is deductively based$_g$ on S's beliefs $Bq_1, \ldots, Bq_n$ only if Bp resulted from S's constructing a class of models for $q_1, \ldots, q_n$ and reading off p as a common conclusion among all the members in that class.

In the simpler cases, there will be only one q_i and the class of models will actually be just a single model that satisfies q_i. It is also possible to have a single model for $n > 1$ premises, depending on the character of the premises. When it comes to belief-support deductive inferential basing (represented by "is deductively based$_s$"), again involving beliefs with symbolic contents, we would have:

> **(BS)** S's belief that p is deductively based$_s$ on S's beliefs $Bq_1, \ldots, Bq_n$ only if Bp was already held by S and S constructed a class of models for $q_1, \ldots, q_n$ and read off p as a common conclusion among all the members of that class.

Notice that both (BG) and (BS) are only *partial* explications of deductive inferential basing: there may be other conditions that need to be satisfied in order for a belief to count as deductively based on other beliefs. They seem however to be instances of *causal* accounts of basing (see Section 1). The causal approach to basing says, roughly, that a belief is based on certain grounds when it is brought about or sustained by means of those grounds. This is particularly evident in (BG), which says that the belief that is inferentially based on the given reasons *resulted from* the subject's construction of models that satisfy the relevant premises and from her activity of reading the conclusion off of those models. It is less evident in (BS), but of course one way to understand it is as saying that the subject's construction of models that satisfy the premises and her reading the conclusion off of those models gives additional causal sustenance to her belief in the conclusion. As such, these explications can be used to precisify or improve upon the general causal notion of basing, which is widely held throughout the literature in contemporary epistemology (see Goldman, 1979; Pollock & Cruz, 1999, just to give some examples). And, of course, (BG) and (BS) are not incompatible with higher order requirements for basing, for each only establishes *a* necessary condition for deductive inferential basing.

(BG) and (BS) are accounts of deductive inferential basing *simpliciter* – they are not yet accounts of deductive basing *in the right way*. In order to derive the latter account, we again have to bring the normative considerations made by the mental model theorist into the picture.

The MM theory says that you reason deductively in the wrong way when you infer that p from premises $q_1, \ldots, q_n$ (because all models you have built for the latter ones are also models for the former one) but fail to consider a model (when there is one) in which both $q_1, \ldots, q_n$ and not-p are satisfied. Let us say that *S* constructs a *representative* class of models for a set of premises $q_1, \ldots, q_n$ with respect to p when and only when either (i) there are no models that satisfy both $q_1, \ldots, q_n$ and not-p and *S* constructs a class of models for $q_1, \ldots, q_n$ that also satisfy p, or (ii) there are models that satisfy both $q_1, \ldots, q_n$ and not-p and *S* constructs at least one model that satisfies both $q_1, \ldots, q_n$ and not-p. The disjunction here is *exclusive*. So, assume that *S* builds a representative class of models for premises $q_1, \ldots, q_n$ with respect to p. Under that assumption, it follows that *if* there are models that satisfy both $q_1, \ldots, q_n$ and not-p, then *S* has built at least one such model.

So we might explicate belief-generation and belief-support deductive inferential basing in the right way, respectively, as follows (for beliefs with symbolic contents):

> **(RG')** S's belief that p is deductively based$_g$ on S's beliefs $Bq_1, \ldots, Bq_n$ *in the right way* only if Bp resulted from S's constructing a *representative* class of models for $q_1, \ldots, q_n$ with respect to p and reading off p as a common conclusion among all the members in that class.

(RS') S's belief that p is deductively based$_s$ on S's beliefs $Bq_1, \ldots, Bq_n$ *in the right way* only if Bp was already held by S and S constructed a *representative* class of models for $q_1, \ldots, q_n$ with respect to p and *S* read off p as a common conclusion among all the members in that class.

These explications stem from a general proposal according to which deductive reasoning in the right way about the truth of p, as related to that of $q_1, \ldots, q_n$, consists in constructing a representative class of models for $q_1, \ldots, q_n$ with respect to p and either reading off p as a common conclusion among the members of that class (when p follows from $q_1, \ldots, q_n$), or finding a model that satisfies both $q_1, \ldots, q_n$ and not-p (when p does not follow from $q_1, \ldots, q_n$).

One immediate complaint about these proposals is that although they may establish necessary conditions for basing in the right way, they are still *too weak* to be sufficiently informative about the nature of basing in the right way. For example, a reasoner may build a model for the premises *All As are Bs* and *x is A* in which the set of *As* is a proper subset of the set of *Bs* (*ergo* in which there are *Bs* that are not *As*) and read off the valid conclusion that *x is B* from that model – but there is a sense in which her cognitive performance was normatively flawed, in that she ignored the possibility of *A* and *B* being coextensive. Although her model was *representative* of the premises, it was not *exhaustive*: there is a certain type of model for the premises that she should have also considered. A new proposal would be, then:

(RG) S's belief that p is deductively based$_g$ on S's beliefs $Bq_1, \ldots, Bq_n$ *in the right way* only if Bp resulted from S's constructing a *representative* and *exhaustive* class of models for $q_1, \ldots, q_n$ with respect to p and reading off p as a common conclusion among all the members in that class.

(RS) S's belief that p is deductively based$_s$ on S's beliefs $Bq_1, \ldots, Bq_n$ *in the right way* only if Bp was already held by *S* and *S* constructed a *representative* and *exhaustive* class of models for $q_1, \ldots, q_n$ with respect to p and read off p as a common conclusion among all the members in that class.

So when the reasoner infers that *x is B* from her beliefs that *All As are Bs* and that *x is A* by only constructing a model in which the set of *As* is a proper subset of the set of *Bs*, she is not quite reasoning in the right way – but she does reason in the right way when she also constructs a model in which *A* and *B* are coextensive. For when the reasoner builds these two types of models for the premises *All As are Bs* and *x is A*, she exhausts the types of models that could satisfy those premises – at least as far as representing *only* facts about *As* and *Bs* is concerned (if all possibilities involving *Cs*, *Ds*, etc. were represented, it would be very hard indeed for the reasoner to exhaust the models that satisfy the premises – we are assuming that a plausible notion of *exhaustion* may be cooked up by the MM-theorist here).

Notice that deductive basing in the right way requires the reasoner to overcome the problems with the *principle of truth* mentioned in Section 3. This is not only because of the representativeness and exhaustiveness criteria for basing in the right way, but because there is a success clause built into the verb *to read off* a conclusion from a class of mental models. To use the example I gave in Section 3 to illustrate how the principle of truth operates in model building: in order for a subject to read the conclusion that *x is a circle if and only if x is not a square* from the two relevant mental models for the premise *Either x is a circle or a square*, she has to recognize that *x* fails to be a square when it is a circle, and that it fails to be a circle when it is a square.

Now, just as (BG) and (BS) may be used to individuate the specifics of a causal construal of inferential basing, so (RG) and (RS) may be used to explicate what it is for a belief to be caused by others *in the right* way in cases of deductive inference. So the defender of a causal account of basing can make full use of these accounts, not only as a means of describing the causal mechanisms behind the production of inferential beliefs, but also as a means of assessing them normatively.

There are many questions to be explored in connection with these accounts of basing. For example, how could we explicate *non-deductive* inferential basing in the right way using similar ideas? Also, can the MM theory say something about basing relation involving primitive, non-symbolic belief? Unfortunately, we have no space to explore these issues in detail here. These questions should be addressed in future work on the mental models theory of inferential basing.

7.

We have fleshed out an account of deductive inferential basing using ideas from the cognitive psychology of reasoning. In particular, we used a theory about *how we reason*: the mental models theory. To the extent that that the MM theory is well-confirmed and fits the data, the account of basing derived from it is empirically informed and empirically motivated. Of course, the MM theory may not turn out to be the best theory about how we reason (or, more modestly, about how we reason in such-and-such circumstances). A full defense of the account of basing presented above would require a full defense of the MM theory itself – something we must leave for future work. We expect this work to foster further exploration on the use of psychological theories to give rise to finer grained and less empirically risky accounts of basing.

Notes

1. For example, see Moser (1989, p. 157). For a causal account of basing in which causation is understood in counterfactual terms, see Swain (Swain, 1981, p. 74). For a more recent causal account of basing that purports to deal with deviant causal chains, see McCain (2012).
2. For a recent proposal according to which the basing relation is a *dependence* relation, see Evans (2013).
3. See also Johnson-Laird, Byrne, and Schaeken (1992), and Johnson-Laird and Khemlani (2013).
4. See Johnson-Laird et al. (1992) for discussion.
5. There is more than one reason why certain possibilities are not represented through mental models. A very general one is that considering too many models overloads our working memories, so we strive for economy. Sometimes issues of relevance also creep in – for example, we may regard the possibilities in which the antecedent of a conditional is false irrelevant, for they would not tell us if there is any positive connection between the truth of the antecedent and the truth of the consequent.
6. See Manktelow (1999, Chapter 3) for an overview of the literature on reasoning with conditionals.
7. The unified mental models theory of reasoning is articulated in Johnson-Laird and Khemlani (2013).
8. We use italicization to represent mentally tokened sentences.
9. This is not to say that the empiricist is hereby trying to give an answer to the question: What does "2 + 2 = 4" *mean*? So far, this is just an approach to the nature of mental models, not an analysis of meaning.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the Alexander von Humboldt-Stiftung.

References

Carnap, R. (1947). *Meaning and necessity*. Chicago, IL: University of Chicago Press.
Evans, I. (2013). The problem of the basing relation. *Synthese, 190*, 2943–2957.
Goldman, A. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and knowledge* (pp. 1–23). Dordrecht: Reidel.
Johnson-Laird, P. N. (1975). Models of deduction. In R. Falmagne (Ed.), *Reasoning: Representation and process* (pp. 7–54). Springdale, NJ: Erlbaum.
Johnson-Laird, P. N. (2008). *How we reason*. New York, NY: Oxford University Press.
Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99*, 418–439.
Johnson-Laird, P. N., & Khemlani, S. (2013). Toward a unified theory of reasoning. *Psychology of Learning and Motivation, 59*, 1–42.
Lewis, D. (1970). General semantics. *Synthese, 22*, 18–67.
Manktelow, K. (1999). *Reasoning and thinking*. East Sussex: Psychology Press.
McCain, K. (2012). The interventionist account of causation and the basing relation. *Philosophical Studies, 159*, 357–382.
Moser, P. (1989). *Knowledge and evidence*. Cambridge: Cambridge University Press.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences, 5*, 349–357.
Pollock, J. L., & Cruz, J. (1999). *Contemporary theories of knowledge* (2nd ed.). Lanham, MD: Rowman & Littlefield.
Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
Swain, M. (1981). *Reasons and knowledge*. Ithaca, NY: Cornell University Press.