The Brain as a Model-Making Machine

Dan Ryder, UNC Chapel Hill

1. Introduction – Representation and function

In this paper, I will introduce you to a new theory of mental representation, emphasizing two important features. First, the theory coheres very well with folk psychology; better, I believe, than its competitors (e.g. Cummins, 1996; Dretske, 1988; Fodor, 1987 and Millikan, 1989, with which it has the most in common), though I will do little by way of direct comparison in this paper. Second, it receives support from current neuroscience. While other theories may be *consistent* with current neuroscience, none that I know of actually receives some degree of *confirmation* from it.

There are many different kinds of representations. Some examples are maps, words, meter and gauge readings, diagrams, pictures, scale models, computer simulations, blueprints, charts, musical notation, smoke signals, semaphore, and computer data structures. Qua representations, they all possess intentionality, or aboutness: maps are about places, most words are about the entities they refer to, meters and gauges are about the quantities they measure, etc. However, it seems they have little in common beyond this aboutness (Millikan, 1984, p. 85). Therefore we should be open to the possibility that the aboutness of different representations is ultimately to be explained in different ways.

It is becoming increasingly popular to understand the aboutness of a large class of these representations in terms of *function*.[1] For example, a tire gauge represents one of the properties that it indicates or carries information about, namely air pressure. However, it also carries information about other quantities. If the pressure and volume of the tire are kept constant, the tire gauge will indicate the temperature of the air inside the tire, and if the temperature and pressure are kept constant, the gauge will indicate the tire volume. However, although the tire gauge indicates these things, it does not represent them. It only

[1] Philosophers who favour a teleosemantic approach include Dretske (1988; 1995), Millikan (1984; 1989), Papineau (1987), Sterelny (1990), van Gulick (1980), Lycan (1996), Jacob (1997), and (partially) Cummins (1996).

represents the tire pressure. The teleological theory of representation in indicators says that the tire gauge represents tire pressure because it is the *function* of the tire gauge to indicate pressure. For a gauge or meter to represent some quantity is for it to have the function of indicating that quantity (Dretske, 1988).

The teleological theory also seems to apply to representation in maps. A map of Oconomowoc, Wisconsin represents Oconomowoc because it has the function of being two-dimensionally similar to it. The map may also happen to be two-dimensionally similar to Blow-me-down, Newfoundland, but it does not represent Blow-me-down because it does not have the *function* of being two-dimensionally similar to Blow-me-down.[2] If it lacks the function of being two-dimensionally similar to anything, a sheet of paper physically indistinguishable from a map of Oconomowoc might represent nothing at all. This could be the case if it forms part of a section of wallpaper, for instance.

The importance of the link between representation and function is that it promises to reveal a way of understanding *mental* representation. The representations I listed earlier – maps, words, gauges, etc. – all seem to be dependent in some way upon intentional use. However, in order to avoid a regress, the aboutness of mental representation *cannot* be dependent upon intentional use. Thus the promise of understanding intentionality as a functional property, since functions *may* be dependent upon intentional use, but they need not be. In particular, it is often argued that evolution by natural selection is capable of endowing items with functional properties (, 1998). Mental representation could well be a functional kind whose aboutness is independent of intentional use.

Note that the functions underlying the aboutness of different types of representation may differ. In gauges and meters, to represent x is to have the function of indicating x, whereas in maps, to represent x is to have the function of being two-dimensionally similar to x. Supposing that mental representation is a genuine, unified *kind* of representation, then, our task is to discover what particular function accounts for the intentionality of mental representations.

[2] Though one might be able imagine circumstances in which it could *acquire* that function.

2. Representation in models

Rather than taking indicators (Dretske, 1988; 1995), or pictures, or words to be the analogue of mental representation, I believe that neuroscience and psychology recommend that we adopt the representational paradigm of *models*.[3] Some of the most familiar examples of models are scale models, like a child's toy model airplane, or a model of a building that is to be constructed. Models capitalize on *isomorphism*. Isomorphism is a relation between two structures (e.g. spatial structures). A structure may be abstractly described as a set of elements that enter into a number of *n*-place relations with each other. Consider a structure S_1, where the elements of S_1 are interrelated by a single type of 2-place relation, R_1, according to some particular pattern. That is, R_1 obtains between certain specific pairs of elements of S_1. S_1 is isomorphic to another structure, S_2, if there is a relation R_2 (also two-place) and a one-to-one function mapping the elements of S_1 onto the elements of S_2 such that: for all x and y belonging to S_1, xR_1y if and only if $f(x)R_2f(y)$. This definition may be extended to n-place relations in the obvious way (Russell, 1927pp. 249-50; Anderson, 1995). Thus, for example, a model airplane is isomorphic to the real plane that it models because there is a mapping between points on their surfaces such that: for every spatial vector *v* that obtains between two points on the surface of the model plane, there is a spatial vector *v*′ obtaining between the corresponding points on the real plane, where (in this case) *v*′ is a multiple of *v*. The plane's spatial structure is "mirrored" in the model.

This mirroring is what makes a model useful. When our access to the thing a model represents is somehow restricted, we can use the model to reason about that thing. For instance, if we did not know what the left wing of the Spirit of St. Louis looked like, we could just consult our model to find out. That is an example of using the model to fill in missing information about the world. Another important use of models is in practical reasoning, in figuring out how to act. For instance, the scale model of a building might be used as a guide for its construction.

[3] The earliest modern proposal of this type is due to Craik (1943).

Scale models are static. In order to accommodate beliefs and desires in a natural way, the models in our heads must be *dynamic* models. The elements of a static model and the isomorphic structure it represents are constants, like the position of the tip of the plane's wing, and the position of the tip of the model's wing. By contrast, in a dynamic model the elements in the isomorphic structures are *variables*. Rather than mirroring spatial structure, a dynamic model mirrors covariational structure. For instance, a model used for weather prediction might have elements that correspond to positions in the atmosphere, where these elements can take on different values depending upon whether there is rain, snow, a hurricane, or clear sky at that position. The values of the elements in the model covary in complex ways, and those covariation relations are meant to mirror covariation relations in the atmosphere.

A weather prediction model is perhaps a little too complex to serve as an illustrative example! The example I will use instead is a model of a sink. A sink consists in a number of different variables, like the radial positions of the knobs, which knob is hot and which cold, the temperature at the tap, and the flow rate at the tap. These variables covary in regular ways as governed by the causal structure of sinks. For instance, if I turn the left knob by a certain number of degrees, the temperature at the tap will change by a corresponding amount in a direction dependent upon whether the left knob is hot or cold. A dynamic model of the sink will have elements that stand in for each of the sink variables, and enter into relations that mirror the covariation relations that obtain in the sink.

The uses of a model that we saw in the case of a static model can also be seen in a dynamic model. Because the relations of covariation in the sink are mirrored in the model of the sink, the dynamic model can be used to fill in missing information, or to guide action. Suppose that for some reason you cannot touch the tap water, and so are ignorant of the water temperature, but that you know the radial positions of the knobs. In order to remedy your ignorance, you could make the knob position stand-ins correspond to the positions of the knobs, and then just read off the water temperature from the stand in for temperature at the tap. Since the model mirrors the covariation relations in the sink, imposing particular values on some of the model's variables

will force the remaining variables to adopt values consistent with the covariational structure of the sink.

This property of dynamic models also allows them to be used in guiding action. For example, suppose that I wish the water coming out of the tap to be at a particular temperature and to flow at a certain rate, but I do not know what positions to put the knobs in to obtain that temperature and flow. I may simply consult my model in order to find this out. I just make the stand-ins for temperature and flow correspond to the values that I want, and then I read off the knob positions from *their* stand-ins in the model. These stand-ins correspond to how the knobs *ought* to be, given my need for that temperature and flow. I can then use this information to turn the knobs to the appropriate positions, and get the temperature and flow that I want.

The reason I have been harping on these two particular uses of models is that they form the basis for an elegant account of judgement and occurrent desire. The propositional attitudes are so-called because they seem to have two separable aspects, the representational aspect, and the attitude aspect. For any proposition, one may take different attitudes towards it – one can judge that it will rain, suppose that it will rain, or hope that it will rain. And for any attitude, one may take that attitude towards many different propositions – one can judge that it will rain, or judge that it will be sunny, or judge that one is in Oconomowoc, Wisconsin. One can explain these facts about psychology by supposing that we harbour propositional representations that may occupy various functional roles. The reason why we can both judge that *p* and desire that *p* for any proposition *p* is that a *single* representation is exercised in both cases. If different propositional attitude types were unrelated properties this attitude systematicity would remain unexplained. Separating the attitudes into two different types of representations, as in Dretske [1988] and Millikan [1989] leaves one vulnerable to this problem.

Model representation, on the other hand, allows us to understand the (occurrent) attitudes as having their representational component in common. Suppose an organism has an internal dynamic model of its environment. When the dynamic model is playing a functional role such that it is sensitive to its environment, and part of the model is made to correspond to the environment by

receiving information from the organism's sensory receptors, the rest of the model will come to correspond to the state of the environment through the process of filling in missing information. This is the functional mode that corresponds to occurrent belief or judgement (and its perceptual equivalent). In this functional mode, the model is supposed to correspond to the current state of the environment.

But we saw that the very same model when put to a different use (i.e. in guiding action) is supposed to correspond to how things *ought* to be, given the needs I have input to the model. Thus if the organism's internal model has built into it how the organism's needs relate to the environment, it would be possible for the organism to feed some basic needs into the model (e.g. basic drive signals), and the model would be caused to correspond to how things ought to be in order to satisfy those needs. (In order to use the model in this way, it would have to be at least partially cut off from the environment, i.e. not be receiving complete information about the current state of the world). In this case, the model will implement the organism's occurrent desires.

No doubt this basic idea needs refinement (for some of this, see Ryder, 2002), but it is a very attractive account of the propositional attitudes. It accounts for attitude systematicity in the way the functionalist proposes, and mental representation emerges as all of a kind. According to this account, mental representation is all *model* representation, and the different attitudes just correspond to different "uses" (i.e. functional occupations) of this basic representation.[4]

Let us return to this basic representation, i.e. representation in models generally speaking. We have not yet analyzed what it is for a model to represent something. Just like representation in indicators and maps, representation in models is a *functional* property. Mere isomorphism is insufficient for representation. A rock outcropping that just happens to be isomorphic to The Spirit of St. Louis does not represent The Spirit of St. Louis. And our dynamic model of the sink may mirror the covariational structure in a number of different systems, like the elevator controls in the CN Tower, or power play plan C of the

[4] Here, I leave it open whether this "use" is to be given a causal role or a teleological reading.

Calgary Flames hockey team.[5] A model represents the structure it has the *function* of mirroring or being isomorphic to. That is why the outcropping does not represent The Spirit of St. Louis, and why our model of the sink is not a model of the elevators in the CN tower, or the Calgary Flames' power play.

Isomorphism or mirroring is a non-representational relation that obtains between structures, which are composed of elements that enter into relations. When two structures are isomorphic, their elements are said to correspond. These two relations, isomorphism and correspondence, are "promoted" to being representational properties when they become functional. A model represents a structure *S* when it has the function of mirroring *S*, and the model's elements then represent the elements of *S* because they have the function of corresponding to them. Thus representation in models comes in two varieties, one for the model, and the other for its elements. A model "models" another structure, while an element of a model "stands in for" an element in another structure. The model of The Spirit of St. Louis models The Spirit of St. Louis, and the left wingtip of the model stands in for the left wingtip of The Spirit of St. Louis.

3. Model production

3.1 Evolution vs. learning

As I mentioned earlier, the standard way of justifying functional ascriptions to items whose functions cannot be derived from intentional use or design is by appeal to evolution. Evolutionary function trades on the paradigm of function in products of design. Natural selection is treated as a designer, "choosing" the good designs because of something that they do, and throwing away the bad designs because they do not measure up. Thus the function of the contractile tissue in the iris is to allow for control of the amount of light entering the eye, because it was made to, i.e. naturally selected for, allowing for control of the amount of light entering the eye.

I have no quibble with the idea that evolution can justify functional ascription. However, any teleological theory of mental representation faces a problem if it relies solely upon natural selection to endow content-determining

[5] Given their dismal record, this would not be surprising.

functions upon brain states. The problem is this: most of our mental representations are not acquired through evolution, but rather through *learning*. Suppose that there are models in the brain. The model that you have of your bedroom is not a model whose isomorphism functions were determined by natural selection! Yet in order for it to be a model *of your bedroom*, it must have the function of mirroring your bedroom. Whence this function? Can a learned model be the product of a process that is, like evolution, analogous to intentional design?

There are two existing accounts of learning that might be co-opted to perform the design role. The first, "neural Darwinism", is supposed to be analogous to evolution by natural selection (Changeux, 1985; Edelman, 1987). However, even supposing the analogy is good enough to support a notion of function, the empirical evidence suggests that a selectional account of brain development is at best radically incomplete (Quartz and Sejnowski, 1997). The other possibility is some sort of reinforcement learning story – the "design" of internal models through reward and punishment. (Dretske [1988] gives an account of belief and belief content that depends upon reinforcement, although it is unclear how Dretske's theory could be extended to the design of internal models.) The problem with this is that not all learning depends upon reinforcement. It seems that new representational capacities can be acquired merely through observation (Bloom, 2000; Goldstone, 1998; Sagi and Tanne, 1994).

My task in this section, then, is to present a theory of internal model acquisition that does not depend entirely upon evolution, nor upon reinforcement, but which is consistent with (or better, supported by) what we know about the brain. Briefly, my story is that our models are designed through a combination of evolution and learning. Evolution designs a *model making machine*, and learning is the operation of this machine.

3.2 Artificial model production

Let us return to models of artifacts, like the model airplane, the weather model, and the model of the sink. How did *these* models come into existence, how did they come to be isomorphic to the things that they model? Typically the

thing that they model serves as a *template* for their production. When someone produces a model of The Spirit of St. Louis, they typically consult the actual plane in producing the model. (Of course, they may do so indirectly through consulting photographs, for example.) When we consider the model produced, and ask the question "What is this a model of?", one way of answering our question would be to tell us what object was used as a template for the model. Since the Spirit of St. Louis was the template for the model produced, it is a model of the Spirit of St. Louis, and not some other plane that it happens to be isomorphic to.

Note how this already begins to move us away from a dependence upon intentional design. Suppose the model designer has an *intention* to produce a model of the Wright biplane, but (mistakenly) uses The Spirit of St. Louis as his template. Is the model that gets produced a model of the Wright biplane, or The Spirit of St. Louis? There are considerations on both sides. We want to eliminate any dependence upon the intentions of a designer. We can do this by *removing* these intentions from the process, that is, by changing to a case of *automated* model production. Consider the following device, "the automatic scale modeler", designed to produce static models. It takes some object as input, and produces a mould from the object. Next it shrinks the mould. Then it injects a substance that hardens inside the mould, and finally it breaks the mould and ejects a small scale model of the original object.

Why is it that we can say that the scale model this machine produces is a model of the original object? Suppose the original object is the Spirit of St. Louis. Note that there need not be any intention to produce a model of the Spirit of St. Louis at work here. Perhaps someone just set this model making machine loose on the world, letting it wander about, making models of whatever it happens to come across. (Of course, there were intentions operative in the production of the machine; what we have eliminated is any specific intention to produce a model *of The Spirit of St. Louis* or anything else.) The scale model produced is a model of the Spirit of St. Louis simply because that plane is the thing that served as a template for production of the model. The function of this machine is not to produce models of particular things. It has the function of making a model *of whatever it is given as input*. Given a model that is produced by such a model

making machine, then, if we want to know what it is a model of, we must know its *history*. The representational content of a model produced by the automatic scale modeler is *broad*. Exactly similar models may be models of different things, if different things served as templates for their production.[6]

In order to figure out the representational content of a particular model, we must also know the machine's design principles. In our example, the spatial structure of the model represents the spatial structure of the thing modeled. But the model has a number of other structural features besides its spatial structure; for example it has a colour similarity structure. However, these other structural features are *not* representational. This is not because these features of the model fail to be isomorphic to the features of the things modeled. It could fortuitously turn out that our scale model of the Spirit of St. Louis has exactly the same colour structure as the Spirit of St. Louis. That doesn't make the colour structure of the model represent the colour structure of the plane.

The colour structure of the model does not represent because if the scale model happened to have a colour similarity structure that mirrored the colour similarity structure of the real plane, this would be entirely by accident, in the sense that *it would not be by design*. That is, the model making machine is not designed to model colour similarity structure. It is only designed to model spatial structure. It is designed so that certain relational features of input objects will cause the production of an isomorphic structure. Those features of the input object that, by design, determine the isomorphism are spatial relations. Thus the model's colour structure does *not* have the function of being isomorphic to the plane's colour structure, but the model's spatial structure *does* have the function of being isomorphic to the plane's spatial structure.[7] Since having the function of being isomorphic is *representation* in models, the model represents the spatial structure of The Spirit of St. Louis, but not its colour structure – even if an isomorphism obtains for both.

---

[6] You may be worried by my description of the model as a model of the Spirit of St. Louis, rather than a model of the type of plane that subsumes the Spirit of St. Louis. Whatever the answer in this case, we will see that for the model-making machine in the brain, there is a determinate answer to the question whether the model is of an individual or of a type, independent of anyone's intentions.

[7] The fact that *spatial* structure models *spatial* structure in our example should not tempt you to think that corresponding relations must always be of the same type.

So the design principles of a model making machine tell us at least two things. First, they tell us, of the many structural features a machine-produced model has, which of these are representational. They also tell us, for each representational structural feature of the model, what type of environmental structure it represents. By itself, isomorphism yields neither of these things. When supplemented with the production history of a particular model, the design principles can tell us exactly what the model and its elements represent, i.e. what the model has the function of mirroring, and what its elements have the function of corresponding to.

Note that the automatic scale modeler is capable of producing *inaccurate* models. Perhaps a piece of the machine falls off during its operation, and introduces a lump into the model of the plane. This model says something false about the plane's structure. Alternatively, it may be that the general design principles for the machine fail in certain unforeseen circumstances, e.g. perhaps deep holes in an object cannot be fully penetrated by the modelling clay. In both of these types of inaccuracies, the machine fails to produce what it is supposed to produce, namely a structure spatially isomorphic to its input.

In the automatic scale modeler, there are two stages to the production of a genuine model with a specific content. The first stage is the intentional design of the model making machine. The second stage is template based production of specific models according to the design principles of the machine. I propose that we can apply these two stages of model production to the brain, in particular to the cerebral cortex (because the thalamocortical system is the most likely brain structure to subserve mentality). The first stage, again, is the design of the model making machine. The cortex, however, is not a product of intentional design, but rather of evolutionary design. The second stage of model production in the cerebral cortex is exactly the same as it is for the scale model: template based production of specific models according to the design principles of the cerebral cortex. This is what it is to acquire new representations through learning.

4. The brain as a model-making machine

The crucial question that now arises is clearly this: what are the design principles of the cerebral cortex? To satisfy the theory of model representation,

the cortex must build dynamic models under the influence of a wide variety of possible environmental templates. A recent theory of learning in the cortex suggests that this is indeed what it does. In this section I will describe, from a functional point of view, the essentials of the SINBAD theory of the cortex in relation to the theory of representation in models I have outlined in the previous sections. For full details of the SINBAD theory, please see (Favorov and Ryder, submitted; Ryder and Favorov, 2001; Ryder, forthcoming).

4.1 Environmental structuring of dynamic models

The SINBAD theory is a theory of cell tuning. A neuron "tunes" to an entity *x* in the environment when it adjusts its connections from other neurons such that it has a strong response to *x* and a weak response to other items (see Fig. 1). The important thing to note is that cell tuning occurs *under the influence of the environment*. I think that we ought to conceive of multiple cells' tuning as a process of template based model production.

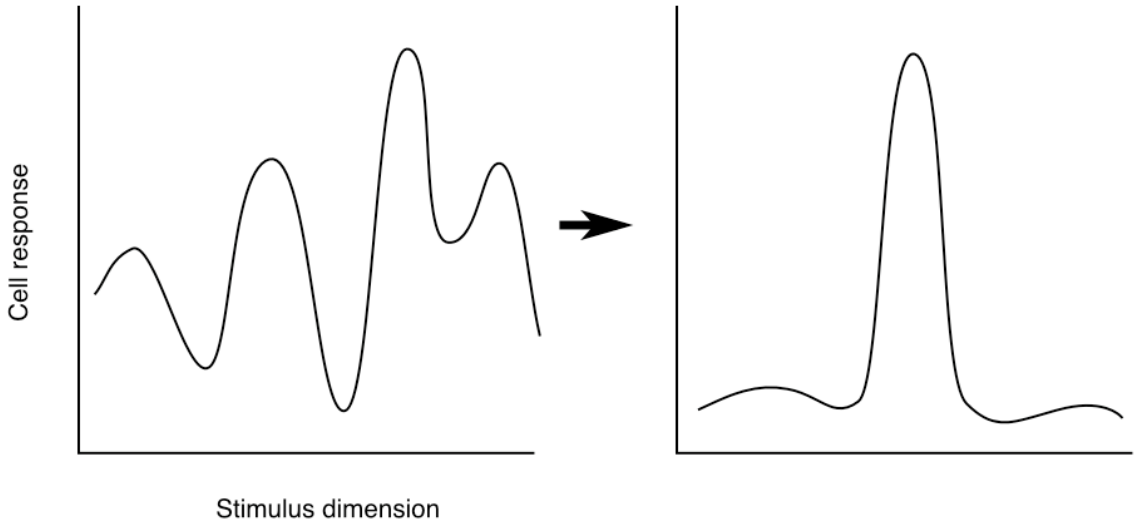

Fig. 1 – Cell tuning

It was important that the automatic scale modeler was designed so that the *represented* structure influenced the production of the *representing* structure in the model. What that means for an automatic dynamic modeler is that the regularity or co-variational structure of the environment must influence the

structuring of the model. A simple example of such dynamic structuring under the influence of the environment would be classical learning by association. The associationist supposes that we begin with internal items that are already "tuned" to particular things in the environment. Taking the neurophysiological point of view, suppose that the internal items are neurons, and that one neuron begins its life tuned to flashes, and another begins its life tuned to booms. Through a process of association, the pairwise correlation between flashes and booms (in thunderstorms) comes to be reflected in a mirroring covariation between the neurons tuned to flashes and booms.

There are a number of reasons why the cortical design principles cannot be those of classical associationism. One particularly serious problem with the associationist proposal is that it cannot, by itself, explain the creation of any genuinely *new* representations (Ryder and Favorov, 2001). The pure associationist requires that we start off with a collection of cells that are *already* tuned to a number of variables in the environment, and the structuring of the model is just a matter of the pairwise correlations amongst the environmental variables coming to be reflected in the correlations amongst the cells already tuned to those variables. On this model, there is no creation of new atomic representations; if we can expand our representational repertoire at all, it can only be through composition.[8] For example, if we start off with a basic set of sensory representations, all of our further representations must be composed of this set's members. The implausibility of this radically empiricist idea (see Fodor, 1998) is what leads the typical associationist to postulate a process of "abstraction" in order to explain our acquisition of concepts of cats and dogs and other things. For our purposes – namely to describe the design principles underlying the cortical model building machine – we would need a mechanistic account of the abstraction process. I will not in fact be taking that route, since I believe that abstraction is an inadequate model of concept formation (Ryder and Favorov, 2001). However, if you disagree, see Grossberg and Carpenter's Adaptive Resonance Theory (Carpenter, 1997; Grossberg, 1976) for an artificial neural network implementation of abstraction that might apply to the cortex.

[8] These initial representations would have indicator functions rather than modeling functions.

There is neurophysiological evidence that the regularity structure in the environment that guides production of cortical models is not simple pairwise correlational structure, as the associationist supposes. Rather, the template regularity pattern is of *multiple* correlations, i.e. multiple features that are all mutually correlated. This proposal is also supported by psychological evidence. While people tend to be quite poor at learning pairwise correlations unless the correlated features are highly salient and the correlation is perfect or near-perfect (Jennings et al., 1982), when multiple mutual correlations are present in a data-set, people suddenly become highly sensitive to covariational structure (Billman and Heit, 1988; Billman and Knutson, 1996; Billman, 1996). This sensitivity to multiple correlations, according to the SINBAD theory, is due to the fundamental architecture of the cortex. It is a consequence of the physiology and anatomy of the principal cortical neuron type, the pyramidal cell.

4.2 Pyramidal cell tuning

A neuron receives inputs on its dendrites, which are the elaborate tree-like structures as depicted on the pyramidal cell in fig. 2. A cortical pyramidal cell typically receives thousands of connections from other neurons, some of which are excitatory, which increase activity, others of which are inhibitory, which decrease activity. (Activity is a generic term for a signal level.) Each principal dendrite – an entire tree-like structure attached to the cell body – produces an activity determined by all of the excitatory and inhibitory inputs that it receives. This activity is that dendrite's *output*, which it passes onto the cell body. The output of the whole cell (which it delivers elsewhere via its axon) is determined in turn by the outputs of its principal dendrites.

The input/output profile of a dendrite, and thus its contribution to the whole cell's output, can be modified by adjusting the strengths of its synaptic connections, and possibly by modifying other properties of the dendrite as well, like its shape (McAllister, 2000; Woolley, 1999). An important question in neuroscience is: What principles underlie the adjustments a cell makes in order to settle on some input to output causal profile? Why do certain connections become highly influential, while others get ignored or even dropped? And what

determines the nature of the influence they come to exert? The SINBAD theory provides one plausible answer to these questions.

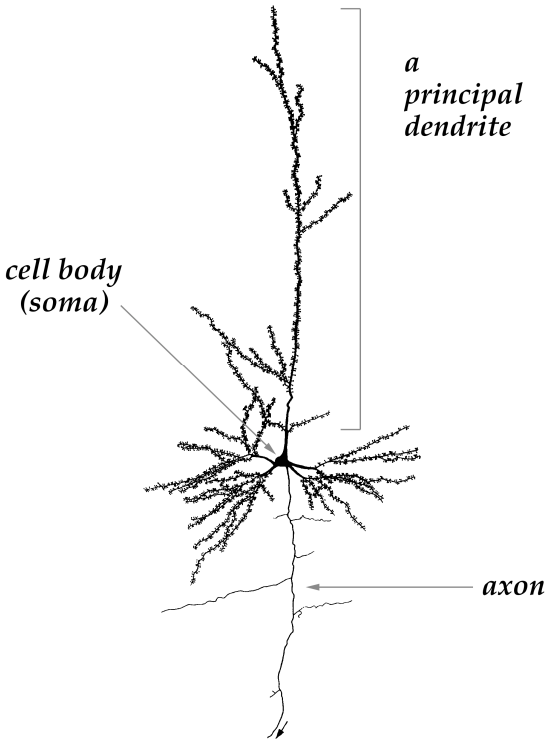

Fig. 2 A typical cortical pyramidal cell. The dendrites form the input region of the cell, which transmits its output via the axon. There are a total of five principal dendrites visible on this cell. ("Dendrite" can refer either to a principal dendrite or a sub-branch of a principal dendrite.) Axons from other neurons synapse on one or more of the thousands of tiny spines covering the dendrites; inhibitory synapses may also occur between spines. This cell type's name describes the pyramid-like shape of the cell body, which is due to the arrangement of principal dendrites (one extending upwards, the others radiating horizontally and obliquely downwards).

The proposal is this: that each principal dendrite will adjust its connections so that it will tend to contribute *the same amount of activity* to the cell's output as the other principal dendrites on the cell. So if there are 5 principal dendrites, like on the cell in fig. 2, they will each tend to adjust their connections over time so that they will consistently contribute $1/5^{th}$ of the cell's total output. I'll put this by saying "They try to match each other's activities." "Trying", of course, is just a convenient metaphor. They are not literally trying, it is merely a brute causal tendency that they have. (The acronym "SINBAD" stands for a Set of INteracting Backpropagating Dendrites, which refers to the mechanism by which the dendrites try to match each other's activities.)

For simplicity, consider a SINBAD cell that has only two principal dendrites. They are trying to contribute an equal amount, 50%, to the cell's

output; that is they are trying to match each other's activities. And they are trying to do that *consistently*, no matter what inputs they happen to get. Suppose the cell's two principal dendrites are connected to the same detector, or sensory receptor. In this situation, it will be very easy for them to match. If they both just pass their input on to the cell body without manipulating it in any way, they will always match.

However, dendrites *do not* get the same inputs, as a rule (Favorov and Kelly, 1996). Thus in the typical situation, the two dendrites' matching task will not be trivial. Suppose, for instance, that they receive two completely unrelated inputs. To use a fanciful example, suppose dendrite *A* receives an input from a green ball detector, while dendrite *B* receives an input from a whistle detector. Suppose both detectors go off at the same time; i.e. there is a green ball present, and there is also a whistle. So both dendrites become active at the same level, let's say 40 units, and they pass that activity on to the cell body, which will become active at 80 units. The dendrites have both passed the *same* amount of activity onto the cell body, so according to the SINBAD connection adjustment principle, they will not change their connections at all. The next time either one receives its input, it will treat it in the same fashion as it did this time.

But remember that it was a *coincidence* that there was a green ball and a whistle present at the same time. Next time, perhaps there is just a green ball. The output of the cell will then be 40 units, where Dendrite *A* accounts for 100% of this output, while dendrite *B* accounts for 0%. The dendrites have radically failed to match. The adjustment principle dictates that Dendrite *A* weaken its connection to the green ball detector, and that Dendrite *B* strengthen any active connections (of which there are none, we are supposing). But it is a hopeless case; the two dendrites will never consistently match activities no matter how they adjust their connection strengths, because they are receiving two utterly unrelated inputs. *The only way they can match is if their inputs are in some way mutually predictable*.

The most basic form of mutual predictability is simple pairwise correlation. If green balls and whistles were consistently correlated, then the two dendrites would be able to match their activities consistently. So, for instance, if Dendrite *A* also received a connection from a beak detector, and Dendrite *B*

received a connection from a feather detector, the dendrites could learn to match. The beak and feather connections would strengthen, while the green ball and whistle connections would weaken to nothing. The learning rule would make Dendrite *A* come to respond strongly to beaks, and Dendrite *B* to feathers. Because beaks and feathers are consistently correlated in the environment, the dendrites will consistently match.

Of course, there are more complex forms of mutual predictability than simple correlation. Real dendrites can receive thousands of inputs, and they are capable of integrating these inputs in complex ways. So the dendrites can find not just simple correlations between beaks and feathers, but also what I call "complex correlations" between *functions* of *multiple* inputs.

Consider another cell. Suppose that amongst the detectors its first dendrite is connected to is a bird detector and a George Washington detector, and for its second dendrite, a roundness detector and a silveriness detector. (Clearly detectors that no well-equipped organism should be without!) There is no consistent simple correlation between any two of these, but there is a consistent *complex* correlation – bird XOR George Washington is correlated with round AND silvery. So in order to consistently match, the dendrites will have to adjust their input/output profiles to satisfy two truth tables. The first dendrite will learn to contribute 50% when [bird XOR George Washington] is satisfied, and the second one will learn to contribute 50% only when [round AND silvery] is satisfied; otherwise they will both be inactive (output = 0). Since these two functions are correlated in the environment, the two dendrites will now always match their activities, and adjustment in this cell will cease.

Consistent environmental correlations are not accidental: there is virtually always a *reason* behind the correlations. For example, the correlation between beaks and feathers in the first example isn't accidental – they are correlated because there exists a natural kind, birds, whose historical nature (an evolutionary lineage) explains why they tend to have both beaks and feathers. What will happen to a cell that has one dendrite that comes to respond to beaks, while the other comes to respond to feathers? The cell will respond to *birds* – the thing that *explains* the correlations in its inputs. Similarly, the second cell will

come to respond to the kind that explains the complex correlations in *its* inputs, namely American quarters.

Mendel postulated the existence of genes in order to explain a set of complex correlations he observed in pea plant crossings. The existence of genes was implicitly suggested by the data that Mendel collected, since it stood to reason that the correlations he observed had some source. What Mendel did was make explicit what was implicitly suggested by his data. Although the "explanatory inference" that a cell mechanizes is significantly less complex, it follows exactly the same pattern. The existence of some source of correlation is suggested by the regularities among the inputs a cell's dendrites receive, and the adjustment principle, which is purely mechanical, makes a cell come to respond to that source, making it explicit.

SINBAD cells thus have a strong tendency to tune to sources of correlation. Different cells will tune to different sources of correlation, depending upon what inputs they receive. Each cell's tuning is to be explained by a particular source, and the correlations that source is responsible for. If tuning to sources of correlation can be understood as a cortical *design principle*, and it results in the production of models that are isomorphic to covariational structures in the environment involving those sources of correlation, then we have our account of template based model production in the brain. Particular sources of correlation, like the kinds *bird* and *quarter*, would be the templates responsible for the creation of elements, namely tuned cells, of an internal model of the environment.

4.3 Cortical SINBAD models

The next thing to show, then, is that pyramidal cell tuning is accompanied by a tendency of cortical networks to structure themselves isomorphically with a large variety of regularities in the environment. Given the obvious utility of such isomorphic structures, this should be enough to make it plausible that it is the *function* of the cortex to build them, i.e. to make it plausible that the cortex is an evolutionarily designed model building machine. Ideally, the modeled regularities should involve sources of correlation that are the sorts of things that *we* represent, if the account is to be a good explanation of mental representation.

First I will consider the sorts of things that fall into the class of sources of correlation. Then we shall see why a SINBAD network has a strong tendency to structure itself isomorphically with regularities involving sources of correlation.

Richard Boyd, Hilary Kornblith, and Ruth Millikan have developed closely related accounts of natural kinds that exhibit them as sources of correlation (Boyd, 1991; Kornblith, 1993; Millikan, 1999). This "unified property cluster" account says that a natural kind is characterized by a set of correlated properties, where some further principle explains *why* they are correlated, and thus why reliable inductive generalizations can be made over them. For example, water is a substance with multiple correlated properties like liquidity in certain conditions, clarity, the ability to dissolve certain other substances, etc., where these "surface properties" are explained by water's nature or hidden essence, namely its molecular structure. This pattern of regularity organized around a source of correlation is not restricted to chemical natural kinds. In the case of biological kinds, these correlations are due, not to an underlying chemical structure, but to the common history shared by their members.

Millikan (1998; 1999) extends the unified property cluster account beyond natural kinds. Non-natural (but real) kinds also have multiple correlated properties unified by some explanatory reason. Artifacts, for instance, will often have correlated properties because they serve some specific function (e.g. screwdrivers), because they originate from the same plan (e.g. Apple's iMac), or because they have been copied for sociological reasons (e.g. a coat of arms and its variants). She also points out that individuals fall into the same pattern. Individuals persist through time, and in doing so tend to retain many of their properties. They are thus sources of a large number of multiple correlations among properties. Events and processes, whether particulars or kinds, are also sources of correlation, for instance Halloween, World War II, biological growth, and atomic fusion. Dynamical systems, like homeostatic systems, economic systems, ecosystems, organisms, and many artifacts, have an underlying causal structure that is a source of correlated variation amongst the systems' global effects. (For example, the flows of water in a sink's incoming pipes explain the complex pattern of correlations observed among the knob positions and the water flow and temperature at the tap.) The point is that this general pattern of

regularities organized around sources of correlation is ubiquitous, although different sorts of unifying principles may underlie different instances of multiple correlations.

Most of the pyramidal cells in the cortex receive most of their inputs from *within* the cortical hierarchy. As a cell's dendrites learn to match activities by satisfying mutually predictive functions of their inputs, the covariational relations among the things the cells tune to come to be mirrored in the causal relations among the cells, mediated by their dendrites. The multiple dendrites on a SINBAD cell must find functions of their inputs that are correlated. Assuming these correlations are not accidental, the cell will tune to their source. In tuning to a source of correlation, a cell will provide neighbouring cells with a useful input, i.e. an input that helps *their* dendrites to find correlated functions. Thus these neighbouring cells, in turn, tune to sources of correlation, and the process repeats. The end result of this complex multiple participant balancing act is that a SINBAD network comes to mirror the regularity structure of the environment from which it receives inputs. It becomes an internal model that reflects the deep structure of the environment, with "stand-ins" for the individuals and kinds around which environmental regularities are structured.

One set of relations between sources of correlation that come to be mirrored in the network are hierarchical relations. For example, early in visual cortex, cells tune to edges, which are sources of correlation responsible for regularities in raw input data.[9] These cells' outputs will be the next cells' inputs, so cells in the next layer will tune to sources of correlation amongst edges, e.g. depth (Becker and Hinton, 1992) and types of translational motion. The next layer may begin to tap into movable shapes (which are sources of correlated motion and edge depths), and so on (Favorov et al., 2002). First there will be a tendency to uncover sources of correlation at progressively larger spatial and temporal scales; then more abstract sources of correlation will be discovered, such as kinds. The nature of the peripheral receptors will clearly exert a large influence on what correlations are available to cells, and thus on what sources of correlation are discovered. Close to the cortical periphery where "raw" sensory

[9] In fact, this "raw" sensory information is highly processed by subcortical areas.

information is received, SINBAD cells will tune to more "subjective" sources of correlation, i.e. ones that depend upon the nature of an organism's receptors. But this subjectivity can be overcome as one proceeds up the cortical hierarchy, and cells tune to objective shapes, kinds, and causes, of increasing orders of generality.[10]

Sources of correlation are also related to each other non-hierarchically. Cats are related to fur and to mice, water is related to taps and salt, and grass is related to greenness and to suburbia. These relations to other sources of correlation form the lion's share of the regularities in which a source of correlation participates. "Lateral connections" in the cortex allow the links between cells to mirror these regularities.

Recall that the usefulness of a model lies in its capacity to fill in missing information, and to guide action. This depends upon a basic process of "filling in" whereby the isomorphism to regularities in the environment is maintained in the face of temporary deviation. "Filling in" is accomplished in a SINBAD network because a cell that has tuned to a particular variable has *multiple* sources from which it can obtain information about that variable, from numerous sensory input channels and also neighbouring cells. If one of those sources of information is blocked off, the others will compensate. It is the fact that SINBAD cells tune to *sources of correlation* that allows them to exhibit this behaviour. The clustering of numerous (possibly complex) properties around a source of correlation allows a cell that tunes to that source to have multiple lines of "evidence" for its presence.[11]

Thus SINBAD networks have a strong tendency to become dynamically isomorphic to regularity structures in the environment, having as their structural elements sources of correlation. Assuming these internal structures are capable

[10] If you are worried that there are far too many sources of correlation our brains need to have some cells tune to, consider the fact that in the densely interconnected human cerebral cortex, there are somewhere between 11 and 25 billion pyramidal cells (Pakkenberg and Gundersen, 1997). (Compare this to a good adult vocabulary of 50,000 words.) If a source of correlation is encountered frequently enough to be predictively relevant, a number of cells are quite likely to get the appropriate inputs to tune to it. There is also a mechanism to prevent too many cells from tuning to the *same* source of correlation – see (Favorov and Ryder, submitted).

[11] A pyramidal cell is not necessarily restricted to 5 to 8 lines of "evidence" from its 5 to 8 principal dendrites. First, multiple functions may overlap on a single dendrite. Second, recent evidence suggests that a cell's separate investigators may not be the principal dendrites, but rather the much larger number of terminal branches (Wei et al., 2001). Third, cells can act in populations, with individual cells in that population tapping into different lines of "evidence" for the presence of the same source of correlation.

of exerting appropriate influence on behaviour (which they are – see Favorov and Ryder, 2001), this was the thing we needed to show in order to make it plausible that the cortex had the *function* of producing dynamic isomorphisms, and thus that it was a model-making machine.

5. Cortical model contents

In keeping with the principles we uncovered in our investigation into template based model production using the example of the automatic scale modeler, these dynamic models in the cortex will have determinate contents insofar as it is possible to identify what regularity structures served as the *template* for their development. Equivalently, we can figure out what items served as the template for production of the *elements* of the model. The elements of a SINBAD model are particular cells that have tuned or are in the process of tuning to a source of correlation.

A SINBAD cell may have causal intercourse with a number of items in the environment, several of which are genuine sources of correlation. However, the cell's template is not just any source of correlation that has helped cause it to fire at some point in its past. We saw that something serves as a template for model production *only relative to the design principles of the model building machine*. So in order to discover what served as a cell's template, and thus what that cell represents, we need to consult the cortical design principles, as described above (if the SINBAD theory is correct).

SINBAD cells are designed to come to correspond to sources of correlation through their dendrites learning to match, where this learning is dependent upon some *particular* source of correlation. This is how the process is supposed to proceed: a cell, which starts off with randomly weighted connections to other cells, is exposed to a source of correlation many times. Upon each exposure, it improves its dendritic matching, finally coming to tune to and correspond to that source of correlation. Now, there can be deviations from the way the tuning is supposed to proceed by design. These will include causal interactions things that have inhibited the cell from achieving its current degree of matching success; or, at least, that have failed to *help* the cell achieve its current matching success. The design principles of the cortical model making machine pick out, as

a cell's template, only the things that have helped that cell achieve matching success. Anything else does not explain the creation of an internal model according to the cortical design principles. Therefore, a single SINBAD cell has the function of corresponding only to the source of correlation that actually helped it achieve the degree of dendritic matching it has attained thus far. *That* is the source of correlation the cell represents. Anything else that it responds to, has responded to, or corresponds to in the context of some isomorphism is not part of the cell's representational content (see Ryder, forthcoming).

There may be several sources of correlation that explain a particular cell's dendritic matching success, such that the cell comes to correpond (to some degree) to several different things. In this case, there is no reason to pick out just one of those sources of correlation as the cell's representational content – the cell is *equivocal*. This is probably the norm. It is made up for, however, by the likelihood that many different cells will come to correspond to the *same* source of correlation, perhaps via slightly different "lines of evidence". Thus a mental representation with a fully determinate content would normally be subserved by a small *population* of cells. (The SINBAD theory is not a theory of "grandmother cells.")

In the perceptual case the network is in the functional mode that characterizes judgement; the model is being "used" to help gather and make explicit information about the state of the environment. In this mode, a SINBAD cell misrepresents when it fires in response to something that is not its template, that does not explain the ability of its dendrites to consistently match activities. It might respond, for instance, to something that has a subset of its template's correlated features. Suppose there is a cell whose dendritic correlations are to be explained by its previous exposure to pennies. Suppose this cell responds to a dime in my pocket. Because it is *pennies*, and not *dimes*, that explain the cell's achievement of (possibly imperfect) dendritic matching, its function is to correspond to pennies, not dimes. Therefore it represents pennies and not dimes. So when it fires in response to a dime, it, or rather, you misrepresent the dime as a penny – the model is being "misused." (This is different from the sort of error that is the construction of an inaccurate model. Misuse is false *judgement*,

inaccuracy is false *belief*.) For further consideration of the sorts of cases that tend to cause trouble for psychosemantic theories, see (Ryder, forthcoming).

This theory of the cortex as a template-based model building device has many virtues. It explains how it is that we can acquire representations through learning. Since cortical models are *dynamic* models, an account of judgement and occurrent desire that upholds folk psychology is made easily available; all that is required is the addition of an account of the functional roles a model representation must occupy in order to implement the various occurrent attitudes. In addition, the theory holds promise as an account of mental representation because the representational contents it predicts are the sorts of contents that characterize *our* mental states – SINBAD representations are about sources of correlation, including natural and other real kinds, as well as individuals. These contents are also *broad*, depending upon a representer's prior interaction with its environment. The theory allows for representational error, including both false belief (inaccurate models) and false judgement (misuse of a model). Last, but not least, it is supported by current neuroscience.

References

Allen, C., Bekoff, M., & Lauder, G. (Eds.). (1998). *Nature's Purposes: Analyses of Function and Design in Biology*. Cambridge, Mass.: MIT Press.

Anderson, C. A. (1995). Isomorphism. In Kim, J., & Sosa, E. (Eds.), *A Companion to Metaphysics*. (pp. 251). Oxford: Blackwell.

Becker, S., & Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*, 161-163.

Billman, D. O. (1996). Structural Biases in Concept Learning: Influences from Multiple Functions. In Medin, D. (Ed.), *The Psychology of Learning and Motivation*. San Diego: Academic Press.

Billman, D. O., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, *12*, 587-625.

Billman, D. O., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 458-475.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, Mass.: MIT Press.

Boyd, R. N. (1991). Realism, anti-foundationalism, and the enthusiasm for natural kinds. *Philosophical Studies*, *61*, 127-148.

Carpenter, G. A. (1997). Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. *Neural Networks*, *10*, 1473-1494.

Changeux, J.-P. (1985). *Neuronal Man: The Biology of Mind*. Oxford: Oxford University Press.

Craik, K. J. (1943). *The Nature of Explanation*. Cambridge: Cambridge University Press.
Cummins, R. (1996). *Representations, Targets, and Attitudes*. Cambridge, Mass.: MIT Press.
Dretske, F. (1988). *Explaining Behavior*. Cambridge, Mass.: MIT Press.
Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, Mass.: MIT Press.
Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.
Favorov, O. V., & Kelly, D. G. (1996). Local receptive field diversity within cortical neuronal populations. In Franzen, O., Johansson, R., & Terenius, L. (Eds.), *Somesthesis and the Neurobiology of the Somatosensory Cortex*. (pp. 395-408). Basel: Birkhauser.
Favorov, O. V., & Ryder, D. (2001). *Learning and making use of nature's order*. Paper presented at the Center for Adaptive Systems, Boston University.
Favorov, O. V., & Ryder, D. (submitted). Neocortical mechanism for discovering hidden environmental variables and their relations: The SINBAD model.
Favorov, O. V., Ryder, D., Hester, J. T., Kelly, D. G., & Tommerdahl, M. (2002). The cortical pyramidal cell as a set of interacting error backpropagating dendrites: a mechanism for discovering nature's order. In Hecht-Nielsen, R., & McKenna, T. (Eds.), *Theories of the Cerebral Cortex*. Berlin: Springer-Verlag.
Fodor, J. (1987). *Psychosemantics*. Cambridge, Mass.: MIT Press.
Fodor, J. (1998). *Concepts: where cognitive science went wrong*. Oxford: Oxford University Press.
Goldstone, R. L. (1998). Perceptual Learning. *Annual Review of Psychology*, *49*, 585-612.
Grossberg, S. (1976). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, *23*, 187-202.
Jacob, P. (1997). *What Minds Can Do*. Cambridge: Cambridge University Press.
Jennings, D. L., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: data-based versus theory-based judgments. In Kahneman, D., Slovic, P., & Tversky, A. (Eds.), *Judgment under Uncertainty: Heuristics and Biases*. (pp. 211-230). Cambridge: Cambridge University Press.
Kornblith, H. (1993). *Inductive Inference and Its Natural Ground*. Cambridge, Mass.: MIT Press.
Lycan, W. G. (1996). *Consciousness and Experience*. Cambridge, Mass.: MIT Press.
McAllister, A. K. (2000). Cellular and molecular mechanisms of dendrite growth. *Cerebral Cortex*, *10*, 963-973.
Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. Cambridge, Mass.: MIT Press.
Millikan, R. (1989). Biosemantics. *Journal of Philosophy*, *86(6)*, 281-297.
Millikan, R. (1998). A common structure for concepts of individuals, stuffs, and real kinds: more Mama, more milk, and more mouse. *BBS*, *21(1)*,
Millikan, R. (1999). Historical kinds and the "special sciences". *Philosophical Studies*, *95(1/2)*, 45-65.
Pakkenberg, B., & Gundersen, H. J. G. (1997). Neocortical neuron number in humans: effect of sex and age. *Journal of Comparitive Neurology*, *384*, 312-320.
Papineau, D. (1987). *Reality and Representation*. Oxford: Blackwell.

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: a constructivist manifesto. *Behavioral and Brain Sciences*, *20*, 537-596.
Russell, B. (1927). *The Analysis of Matter*. London: Routledge.
Ryder, D. (2002) *Neurosemantics: a theory*. Unpublished University of North Carolina: Chapel Hill.
Ryder, D. (forthcoming). SINBAD Neurosemantics: A theory of mental representation. *Mind & Language*,
Ryder, D., & Favorov, O. V. (2001). The New Associationism: A neural explanation for the predictive powers of cerebral cortex. *Brain & Mind*, *2(2)*, 161-194.
Sagi, D., & Tanne, D. (1994). Perceptual learning: learning to see. *Current Opinion in Neurobiology*, *4*, 195-199.
Sterelny, K. (1990). *The Representational Theory of Mind*. Oxford: Blackwell.
Wei, D.-S., Mei, Y.-A., Bagal, A., Kao, J. P. Y., Thompson, S. M., & Tang, C.-M. (2001). Compartmentalized and binary behavior of terminal dendrites in hippocampal pyramidal neurons. *Science*, *293*, 2272-2275.
Woolley, C. S. (1999). Structural plasticity of dendrites. In Stuart, G., Spruston, N., & Häusser, M. (Eds.), *Dendrites*. (pp. 339-364). Oxford: Oxford University Press.
van Gulick, R. (1980). Functionalism, Information, and Content. *Nature and System*, *2*, 139-162.