Published in New Essays in Teleosemantics, *Graham Macdonald and David Papineau, eds., Oxford University Press, 2006. Penultimate draft; please cite published version.*

On thinking of kinds: a neuroscientific perspective

Dan Ryder, UBC Okanagan

1. Kinds and psychosemantics – not a match made in heaven

Reductive, naturalistic psychosemantic theories do not have a good track record when it comes to accommodating the representation of kinds. In this paper, I will suggest a particular teleosemantic strategy to solve this problem, grounded in the neurocomputational details of the cerebral cortex. It is a strategy with some parallels to one that Ruth Millikan has suggested, but to which insufficient attention has been paid. This lack of attention is perhaps due to a lack of appreciation for the severity of the problem, so I begin by explaining why the situation is indeed a dire one.

One of the main tasks for a naturalistic psychosemantic theory is to describe how the extensions of mental representations are determined. (Such a theory may also attempt to account for other aspects of the "meaning" of mental representations, if there are any.) Some mental representations, e.g. the concept of water, denote kinds (I shall be assuming this is non-negotiable). How is this possible? Unfortunately, I haven't the space to canvass all the theories out there and show that each one fails to accommodate the representation of kinds, but I will point out the major types of problems that arise for the kinds of theories that, judging by the literature, are considered viable contenders.[1] In general, the theories either attempt and fail to account for the representation of kinds, or they fall back on something like an intention to refer to a kind – not exactly the most auspicious move for a reductive theory.

There are a number of problems that prevent non-teleosemantic theories from explaining how it is possible to represent kinds. A concept of a kind K must exclude from its extension things that superficially resemble K but whose underlying nature is different, e.g a concept of water excludes XYZ (Putnam, 1975). Any psychosemantic theory that depends exclusively upon the intrinsic properties (including dispositions) of the representer to determine extension will thus fail to provide for the unequivocal representation of kinds, by reason of the familiar twin cases. This problem infects theories based on isomorphism (Cummins, 1996), information (Usher, 2001), and nomological covariance, including (as Aydede [1997] demonstrates) Fodor's asymmetric dependence theory in its latest guise (Fodor, 1994).

---

[1] For example, I ignore pure internal conceptual role theories, since they fail to explain how a concept can even *have* an extension. No psychosemantics has ever been *given* for two-factor theories (e.g. Block, 1986) or "long-armed" role theories (e.g. Harman, 1987), and the early informational theories (Stampe, 1977; Dretske, 1981) have been shown to suffer fatal problems with disjunctivitis and the like (Fodor, 1990).

Some information-based and nomological covariance theories avoid the twin problem by adding further conditions on extension determination. Whether or not these moves help with twins, they do not help with kinds. For example, Prinz 2000; 2002) requires a representation's content to be its "incipient cause", the thing that explains the concept's acquisition. Such modified informational and nomological covariance theories inevitably fall victim to a second problem for such theories, the problem of epistemically ideal conditions. On a nomological covariance theory, surely it is the category that exhibits the *best* nomological covariance with a representation that is its content. (What else could it be?) Since the covariation of a representation with its content will always be better in ideal epistemic circumstances, a nomological covariance theory will always incorrectly dictate that the content of a representation of kind K is rather K-in-epistemically-ideal-circumstances.

Getting the right content cannot be achieved by ruling out those factors that are "merely epistemic," e.g. by adding an "ideal conditions" clause (Stampe, 1977; Stalnaker, 1984), since which factors are merely epistemic will depend upon which factors are semantic (McLaughlin, 1987). If the symbol really means K-in-good-light, being in good light is not a merely epistemic factor. (And if the symbol really means K-in-weak-light, good light will *not* be ideal). The only remaining way to rule out the epistemic factors from the content of a representation on a nomological covariance theory would be by *ad hoc* restriction of the class of candidate contents to those that fit with how we in fact carve up the world. Because we find these categories intuitive, this move can easily go unnoticed. But this is to ignore the fact that a psychosemantics will be part of an *explanation* for why those categories are the intuitive ones, i.e. why our representations have the contents they do. The restrictions on content must be *predicted* by a complete psychosemantics; they cannot be tacked on in order to fix a flawed theory.

An appropriate restriction might *appear* to emerge naturally from nomological covariance theories, since on such theories, the classes of representable contents are presumably marked by having "nomic" rather than "non-nomic" properties in common. But making that distinction principled while ruling out the problem contents and ruling in the legitimate contents is a tall order (not to mention the collateral problem that arises for representing individuals). For instance, if *crumpled*, *shirtness*, and even *crumpled-shirtness* are nomic properties (Fodor, 1991), why on earth isn't gold-in-good-light? (Thus, in Dennett's Philosophical Lexicon: jerry-mander, v. To tailor one's metaphysics so as to produce results convenient for the philosophy of mind.) On the other hand, if we were to restrict representable contents to scientific kinds, the content of much of our mental life would go unaccounted for.

One might expect teleosemantic theories to do better, but in fact they do not. Consider indicator teleosemantics, for instance (Dretske, 1986; Dretske, 1988; Matthen, 1988). If an indicator

I is selected for indicating kind **K**, that means **I**'s indicating kind **K** has been causally relevant to the presence of **I** in this (species of) organism, or causally relevant to **I**'s recruitment to perform some biologically relevant task. The problem that arises is a version of Fodor's (1990) indeterminacy complaint (which is itself a version of the disjunction problem): why suppose that **I** has been selected to indicate **K** as opposed to a disjunction of local signs of **K**? If **K** has been causally relevant to **I**'s selection, then the local signs of **K** that mediate **I**'s ability to indicate **K** have also been causally relevant to **I**'s selection. There is no reason to pick **K** rather than local signs of **K** as **I**'s content, and, again, an *ad hoc* restriction isn't kosher. (In correspondence, Dretske has agreed that he faces difficulties in accounting for the representation of kinds, and, more seriously, anything distal.)

A similar problem infects the basic version of Millikan's "mapping" theory, which involves the selectional recruitment of "representation producers" to produce representations that map onto the environment according to a certain rule. This selectional recruitment operates via these representations' consumers; it is the mapping of the representations onto their contents that has, historically, explained the proper performance of one or more of the consumers' functions – in Millikan's terminology, this mapping is a Normal condition on proper performance of the consumers' functions. She also requires the mapping be something that the producer can bring about or effect, e.g. via a detection mechanism.

Now, whenever Millikan is tempted to suppose that it is a *kind* that a representation is supposed to map onto, we may counter as follows: that kind will have a cluster of typical properties, some subset of which will be those that explain its detection (or however the mapping is brought about) and the proper performance of the consumer's relevant functions. Why not say that the representation denotes either the subset or disjunction of selectionally explanatory properties, rather than the kind itself? A kind is not identical to a subset or disjunction of its properties. This seems to raise the spectre of twin problems for Millikan –the selectionally explanatory properties could be clarity, liquidity, and potability, properties that H_2O and XYZ share. (We shall see later that some additional resources allow Millikan to deal with this problem, if not for the ubiquitous fly-snapping frog, then at least for *us*.)

As far as I know, all other reductive theories face a relative of one of the problems above. We must face the possibility that a reductive naturalistic psychosemantics cannot explain the representation of kinds directly. However, there are also a number of *non*-reductive strategies for explaining the possibility of representing kinds, strategies that appeal in some way to the representation wielder's intentions. Such an appeal might be used in a reductionist project if the representations (normally concepts) mobilized in these intentions do not themselves include any representations of kinds. I think this hope is forlorn, but the reductionist has a lesson to learn from the attempt.

What sort of intention would do the trick? Well, an intention to pick out a kind, of course. For instance, the representation of specific kinds could be accounted for by mental description. Just as one might say that unicorns are picked out in thought descriptionally ("a horse with a horn" – otherwise a puzzling case, especially for the naturalist, since unicorns do not exist), perhaps one could pick out a specific kind with "a ϕ that is a kind", where "ϕ" denotes some complex property in accordance with your favourite reductive theory. Now, I have here characterized the intention to pick out a kind using the concept of kindhood – not exactly a propitious move for the reductionist. If concepts of *particular* kinds are difficult to account for, then the concept of *kindhood* seems even more difficult. And how would such a concept be acquired without prior concepts of specific kinds as examples?

Perhaps a further non-reductive move could be made, filling out the concept of kindhood descriptionally. A well-developed account of kindhood exists that could be put to use in this way. According to the "unified property cluster" account (Boyd, 1991; Kornblith, 1993; Millikan, 1999; Millikan, 2000), 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" (as I call it) 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; 2000; see also Bloom, 2000) 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). Even kinds of events and processes are sources of correlation, for instance Halloween, biological growth, and atomic fusion. (Millikan also points out that individuals fall into the same pattern, although we will not be concerned with them here.)

Now the non-reductionist account of a kind concept becomes "a ϕ that is a member of a class whose syndrome of properties shares a common underlying explanation" or something like that, presumably including in ϕ or mentioning elsewhere in the description some specific features of the syndrome (and perhaps that it is "around here"). This seems to be the idea behind the theory of "psychological essentialism" (Keil, 1989; Medin and Ortony, 1989), and at least one philosopher has appealed to it in order to explain the possibility of representing kinds (Prinz, 2002). Again,

though, this seems cold comfort to the reductionist, for at least two reasons. First, the sophistication of this conception is such that the potential for reduction is not getting any clearer (not to mention the fact that it may put kind concepts beyond the reach of a sizeable proportion of the population, and I don't just mean children). Second, it seems to betray an almost classical empiricist faith that the concepts featured in a kind's syndrome can ultimately be understood (presumably through many steps of analysis), not as kind concepts themselves, but as some other sort of concept. (Perhaps as representations of properties transduced by perceptual systems? This is the extreme view that Prinz's theory entails, although he does not assert it.) This seems unlikely, especially for biological kinds. For example, maleness is a biological kind having a biological syndrome including, e.g. produces sperm, which is its own biological kind, etc.

Doubtless there is much more to be said about the descriptionist strategy. For instance, there is the "division of linguistic labour" (Putnam, 1975), whereby people can supposedly think of kinds through language, by deference to experts on the referents of public language terms. (But how do the experts think of kinds? And does this deference not involve concepts of kinds, e.g. the kind *word*?) Or the description may be "naturalized" by giving it a causal role reading (but how can causal roles determine extensions?) I will not pursue this line of inquiry any further here; for one thing, there are many independent problems that accrue to descriptional accounts of concepts generally (Fodor, 1998; Millikan, 2000). A direct reductionist approach is needed, or at least would clearly be welcome.

Teleosemantics is the reductionist approach I advocate, and the teleosemanticist can learn from the non-reductive appeal to further intentions. The standard teleosemantic strategy is to take what might appear to be *intentions*, an agent's purposes, and reconstrue them as *functions*, or biological purposes. This, I suggest, is what we should do with psychological essentialism. Appeals to intentions to denote kinds lead to murky waters. What we need is for the mind's representational system, or a part of it, to have the *function* of indicating kinds (Dretskean formulation), or that it be *supposed* to map onto kinds (Millikanian formulation), or something similar. That is, kindhood *itself*, as characterized by the unified property cluster account, needs to be selectionally relevant to the representational system.

This, in fact, is a way of understanding what Millikan has attempted over the course of a number of publications (1984 part IV; 1998; 2000), except that she claims selectional relevance not for kindhood, but something broader that includes kindhood as a subtype (she calls it substancehood). The account I present below was developed independently, but has much in common with Millikan's. What is particularly interesting is that my story comes from the neuroscience, whereas hers is derived from abstract psychological considerations. This convergence is surely a sign of truth! The main idea underlying my proposal is that the brain's predictive network

was selected for because of the way it interacts specifically with kinds (actually, the more general class of "sources of correlation"). Its predictive capacities are dependent upon its interacting with kinds *qua* kinds, so kindhood, as characterized in the unified property cluster account, was selectionally relevant to the design of our representational systems. But getting to this conclusion will require considerable stage-setting. First I will present a particular teleosemantic framework, and apply it to representation acquisition (for concepts of kinds are *acquired*); then I will outline the neuroscientific details that yield an answer to our main question.

2. The teleosemantics of models

Many representations may be understood in teleosemantic terms. Although a tire gauge carries information about pressure, temperature, and volume, it represents only pressure because that is what it is *supposed* to carry information about (Dretske, 1986; Dretske, 1988). A map of Oconomowoc, Wisconsin is two-dimensionally similar to both Oconomowoc and Blow-me-down, Newfoundland, but it only represents Oconomowoc because that is the location to which it is *supposed* to be two-dimensionally similar. (And a sheet of paper physically indistinguishable from a map of Oconomowoc that is not a map but rather a section of wallpaper is not supposed to be two-dimensionally similar to *anything*, and as a result, does not represent anything.) Bar graphs, pictures, and many other representations can be treated analogously. For all of these representations, there is some type of relation that they tend to enter into with the things they represent, and *which* thing they represent appears to be the thing with which they are *supposed* to be so related, or to which they have the *function* of being so related.[2]

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*. 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), where a mapping of elements from one structure to the other preserves some pattern of relations across the mapping.[3] This mirroring of a pattern of relations is what makes a model useful.

---

[2] Teleological theories typically put this "supposed to" in terms of *function*. Here, this stretches the normal use of "function" a little; normally we say that something has the function of *doing* something, not of *being* a certain way. However, it is convenient to use the term to cover both sorts of supposed-tos, and this is how I shall use it. What matters is the normativity, not the functionality per se.

[3] 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

When our access to the thing a model represents is somehow restricted, we can use the model to reason about that thing (Swoyer, 1991; Cummins, 1996). For instance, if we do not know what the left wing of the Spirit of St. Louis looks like, we can just consult our model to find out. That is an example of using the model to fill in missing information about the world ("predictive use", broadly speaking). Another important use of models is in practical reasoning, in figuring out how to act ("directive use"). For instance, the scale model of a building might be used as a guide for its construction. (Elsewhere, I have argued that the occurrent attitudes are the causal role equivalents of these two uses [2002].)

Just like representation in indicators and maps, representation in models is a functional property - mere isomorphism is insufficient. A rock outcropping that just happens to be isomorphic to The Spirit of St. Louis does not represent The Spirit of St. Louis because the isomorphism in question is not a normative one – the rock is not *supposed* to be isomorphic to The Spirit of St. Louis. A model represents because it has the *function* of mirroring or being isomorphic to some other structure.[4]

Structures are composed of elements that enter into relations. When two structures are isomorphic, an element of one is said to *correspond* to a particular element in the other, within the context of that isomorphism. These two relations, isomorphism and correspondence, are promoted to being representational properties when they become normative/functional. A model represents a structure *S* when it has the function of being isomorphic to *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 related varieties, one for the model, and the other for its elements. A model of The Spirit of St. Louis *models* The Spirit of St. Louis, while the left wingtip of the model *stands in for* the left wingtip of The Spirit of St. Louis.

3. Model building

Any teleological theory of mental representation faces a problem if it relies solely upon natural selection to endow content-determining functions upon brain states. The problem is this: most of our mental representations, including our representations of kinds, are not acquired through evolution, but rather through *learning* (pace Fodor, 1981; Fodor, 1998). Suppose that there are models in the brain. The internal model that you have of the rules of chess is not a model whose

---

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, 1927, pp. 249-50; Anderson, 1995).

[4] Actually, normative isomorphism isn't sufficient for representation – the representing structure must also be actually or normatively put to one or more characteristic uses, e.g. predictive and/or directive use. See Ryder (2002).

isomorphism functions were determined by natural selection! Yet in order for it to be a model *of the rules of chess*, it must have the function of mirroring the rules of chess. Whence this function?

One teleosemantic strategy for answering this question is to treat learning as a selectional process itself, leading to the acquisition of functions (Papineau, 1987; Dretske, 1988; Papineau, 1993). There are two types of selectional processes that might be co-opted to perform the design role. At the neurophysiological level, there is, "neural Darwinism", where neurons die off in a way that is supposed to be analogous to 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 and learning is at best radically incomplete (Quartz and Sejnowski, 1997). The other possibility, at the psychological level this time, is some sort of reinforcement learning story – "design" through reward and punishment. (Both Dretske and Papineau rely on this.) The problem here is that new representational capacities can be acquired merely through observation, independently of reinforcement (Sagi and Tanne, 1994; Bloom, 2000).

In this section, I describe a framework that yields an account of non-selectional learning in the form of model acquisition. In the next section, I apply this framework to the brain. Briefly, my story is that evolution designs a *model making machine*, and the operation of this machine constitutes learning. (The general framework is closely related to and probably an instance of Millikan's account of "derived relational proper functions". For the purposes of my argument here, though, I prefer to leave it open whether Millikan's more general story is correct.)

I will start with the *intentional* design of models, like the model airplane. 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 by 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, which is necessary if we eventually want to apply it to learning. 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 can *further* reduce the involvement of intentions by moving 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. (It is a big machine!) 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.) The scale model produced is a model of the Spirit of St. Louis simply because the plane is what served as a template for production of the model. Clearly, the function of this machine is not to produce isomorphs of particular things. It has the more general function of producing isomorphs *of whatever it is given as input*. Each individual model inherits its function of mirroring some specific object **O** from this general function, and the fact that **O** is the input that figures in its causal history. Consequently, for any particular model them machine produces, we must know that model's causal history in order to know what it represents.

But there is something else we need to 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 density structure. However, these other structural features are not representational. Even if it fortuitously turned out that our scale model of the Spirit of St. Louis has exactly the same density structure as the Spirit of St. Louis, the density structure of the model would not correctly represent the density structure of the plane (just as a black and white TV doesn't correctly represent the colour of a zebra). This is because if the scale model happened to have a density structure that mirrored the density structure of the real plane, it would be entirely by accident, in the sense that it would not be *by design*.

A model making machine is designed so that certain specific types of relational features of input objects will cause the production of a specific type of isomorphic structure. Those features of the input object that, *by design*, determine the isomorphism for the automatic scale modeler are spatial relations – and so spatial relations are the only relations the model represents, that it has the function of mirroring. Similarly, the only relational features of the model that are structured by the input object, *by design*, are spatial relations. Thus only the spatial features of the model represent. 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 being isomorphic to, and what its elements have the function of corresponding to in the context of that isomorphism.

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. 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 is the design of the model making machine, either intentional design (the automatic scale modeler) or evolutionary design (the cortex). The second stage is exactly the same in both: template based production of specific models according to the design principles of the machine, as determined by the first stage. This is what it is to acquire new representations through (non-reinforcement) learning.

If we suppose that the seat of the mind, the cerebral cortex, is designed (by natural selection) to build models of the environment, the crucial question that arises is this: what are the design principles of the cortex? In the next section I will describe, from a functional point of view, the essentials of these design principles according to the SINBAD theory. First, though, a little preview of how this foray into neuroscience will help us eventually answer the question we started out with, of how it is possible to represent kinds.

The type of models the cortex is designed to build are *dynamic* models.[5,] 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 likely to be 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. (Weather models are used only predictively, to fill in missing information [about the future]; but if we had the ability to manipulate some weather-affecting variables, such models could be given directive use as well.)

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

[5] The earliest extended physicalist discussion of the dynamic isomorphism idea, and a defense of its relation to the mind, occurs in Kenneth Craik's *The Nature of Explanation* (1943). See also Cummins (1989) and McGinn (1989, Ch. 3).

occurs *under the influence of the environment*. I think that we ought to conceive of multiple cells' tuning as a process of template based production of dynamic models.

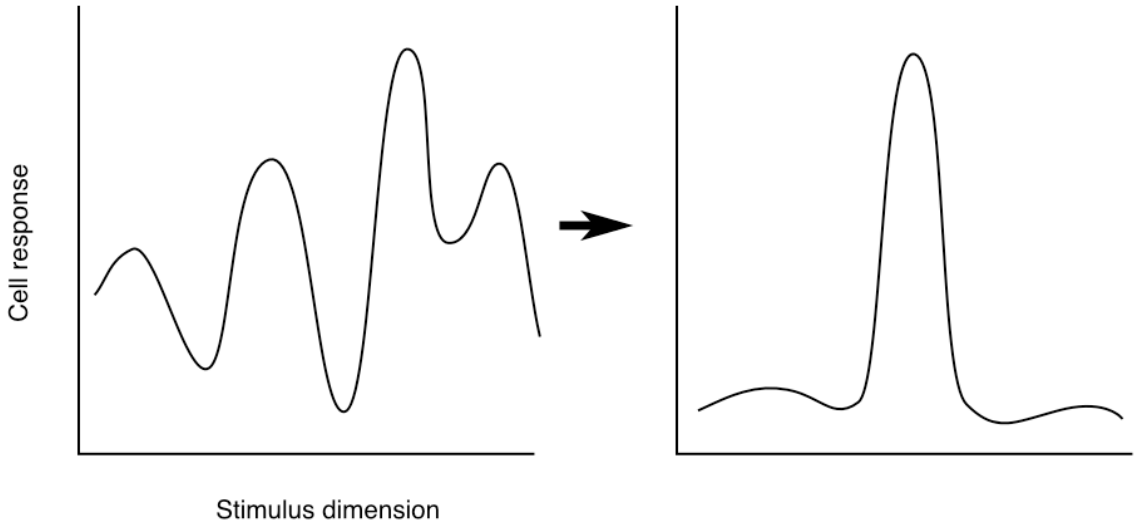

Fig. 1 – Cell tuning

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 (fig. 2a). 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 is too impoverished to explain our capacity to reason (Fodor, 1983). In any case, 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 (fig. 2b) (Favorov and Ryder, forthcoming). This proposal also receives support from psychology. 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).

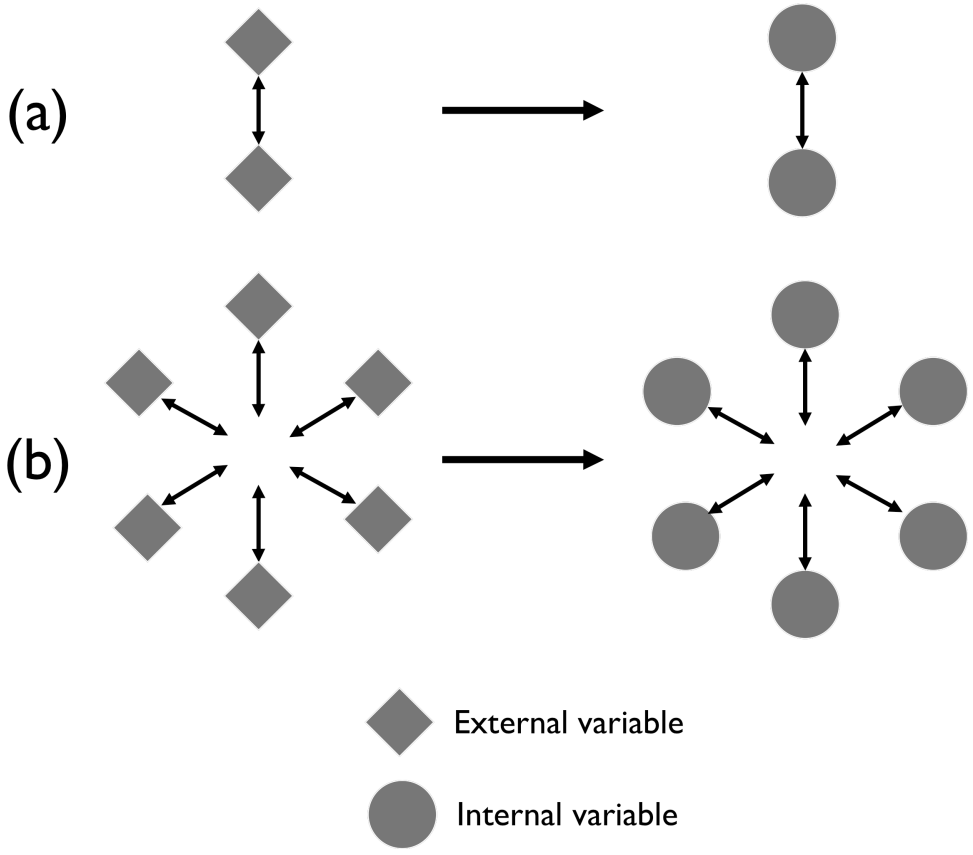

Fig. 2 - In (a), mirroring of pairwise correlations (associationism); in (b), mirroring of multiple correlations.

Already, a special relationship between the cortex and kinds is intimated. Recall the unified property cluster account of kinds; it said that a kind is characterized by a set of correlated properties, i.e. multiple mutual correlations, where some further principle explains why they are correlated. According to the SINBAD theory, the principal cells of the cerebral cortex are built to take advantage of this general pattern of regularity, the pattern due to sources of correlation (including kinds). Interacting with sources of correlation allows SINBAD networks to become dynamically isomorphic to the environment, making them useful for prediction (and direction).

4. The SINBAD theory of the cerebral cortex[6]

The relevant cortical design principles apply in the first instance to *pyramidal cells* (see fig. 3), the most common neuron type in the cerebral cortex (70% to 80% of the neurons in the cortex fall into this class - see Abeles, 1991; Douglas and Martin, 1998). Like any other neuron, a pyramidal cell receives inputs on its dendrites, which are the elaborate tree-like structures as depicted on the cell in fig. 3. A cortical pyramidal cell typically receives thousands of connections from other neurons, some of which are excitatory, which increase activity, and 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.

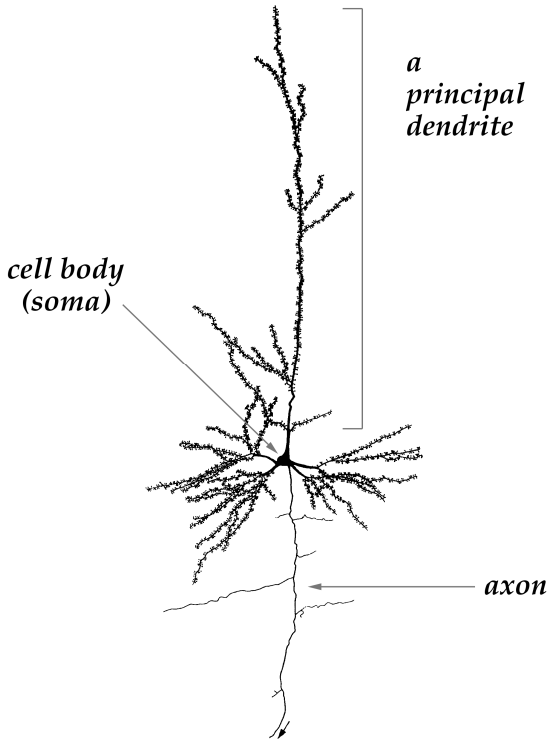

Fig. 3 A typical cortical pyramidal cell. The dendrites form the input region of the cell, which transmits its output via the axon. There are a total of five principal dendrites visible on this cell. ("Dendrite" can refer either to a principal dendrite or a sub-branch of a principal dendrite.) Axons from other neurons synapse on one or more of the thousands of tiny spines covering the dendrites; inhibitory synapses may also occur between spines.

[6] For full details of the SINBAD theory, please see Ryder and Favorov (2001), Ryder (forthcoming), and Favorov and Ryder (forthcoming)(Favorov and Ryder, forthcoming).

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 (Woolley, 1999; McAllister, 2000). 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? For short: what is the pyramidal cell "learning rule"? The SINBAD theory provides one plausible answer.

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. 3, they will each tend to adjust their connections over time so that they will consistently contribute 1/5th of the cell's total output. I'll put this by saying "They try to match each other's activities." They are not literally trying, of course, 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 also a whistle sounds. 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 on to 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.

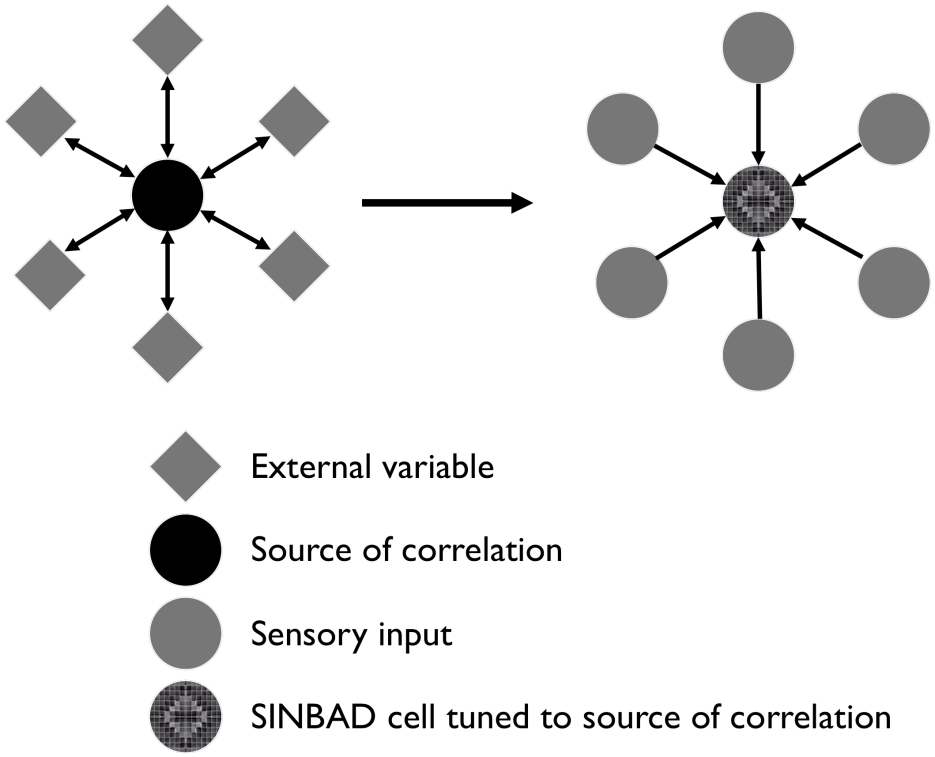

Fig. 4 – A SINBAD cell tunes to a source of correlation by selecting sensory inputs from the mutually correlated external variables that constitute that source of correlation's syndrome of features.

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 (fig. 4). When this tuning process takes place over an entire network, the network is transformed so that its flows of activation come to mirror regular variation in its containing organism's environment. Where the environment has some important variable – a source of correlation – the network will have a cell that has tuned to that source of correlation. And where there is a predictive relation among sources of correlation, the network will be disposed to *mirror* that relation. The activities of the network's cells will covary in just the way that their correspondents in the environment do. In short, the network becomes dynamically isomorphic to the environment.

The reason that a cortical SINBAD network develops into a dynamic isomorphism is that cells' inputs are not only sensory, but are also (in fact primarily) derived from within the cortical network. A cell's tuning is guided, in part, by these intracortical connections (Phillips and Singer, 1997). Tuning – changing a cell's dispositions to react to the environment – occurs through the

modification of a cell's dispositions to react to activity in other cells, mediated by intracortical connections. It is these latter dispositions that come to mirror environmental regularities.

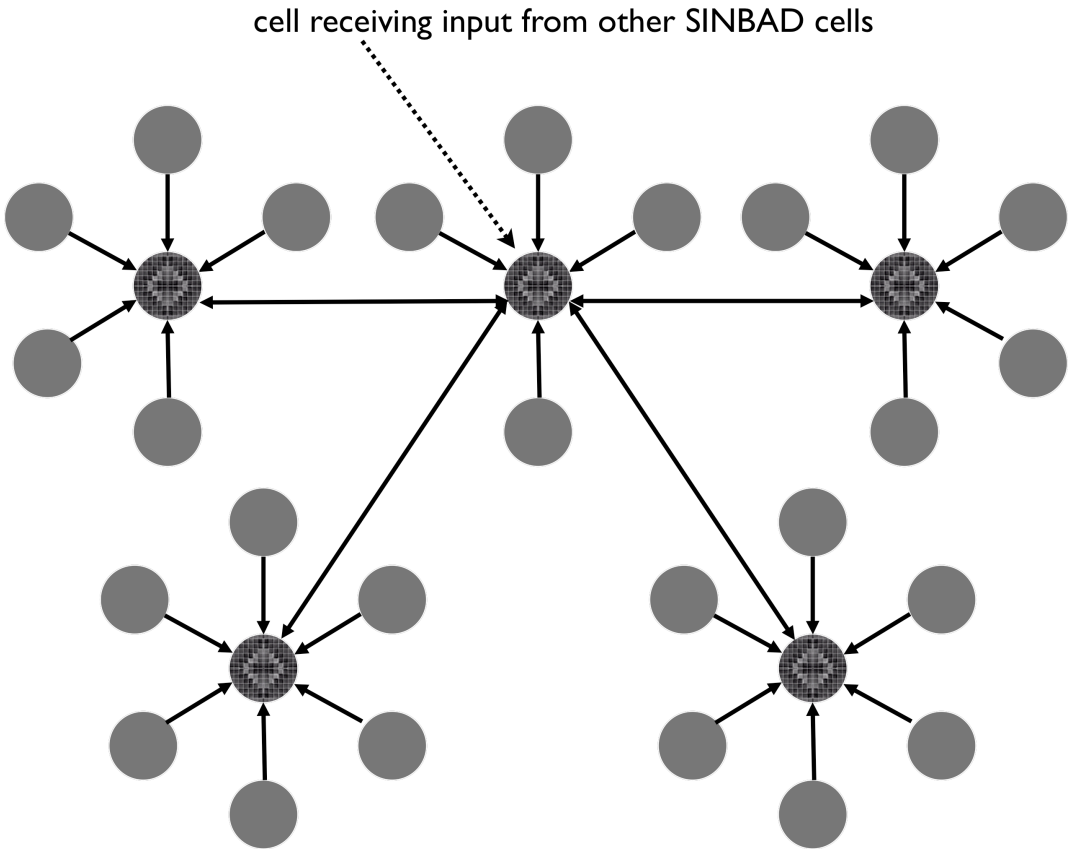

Fig. 5 – When a SINBAD cell receives intracortical inputs from other SINBAD cells, the relations among sources of correlation come to be mirrored in those connections. (The mirroring is shown as bidirectional since cortical connections tend to be reciprocal, though the two directions are mediated by distinct "cables".)

Remember on the classical associationist picture, a pairwise correlation in the environment comes to be mirrored in the brain (fig. 2a). On the SINBAD picture, it's not simple pairwise correlation that comes to be mirrored in the brain, but patterns of multiple correlations, plus their sources (fig. 4). In fig. 5, where one SINBAD cell's inputs come from other SINBAD cells in the cortical network, it can be seen that, through the process of cell tuning, the regularities obtaining among the sources of correlation upon which these cells depend for dendritic matching will come to be reflected in their intracortical connections. Since sources of correlation are interrelated both within and across levels (cats are related to fur and to mice, water is related to taps and salt, and grass is related to greenness and to suburbia), an extensive network develops, as a cell's dendrites come to use other cells' outputs in finding a function that allows them to match. A cell may start

with a tenuous correlational seed[7], but this subtle sign of those correlations' source is enough to put the cell on a path towards discovering the multitude of regularities in which that source participates. As the cell achieves more and more robust dendritic matching, the correlational seed ends up producing a complex dendritic tree, which realizes complex functions relating that source of correlation to many others (or rather cells that have tuned to them).

Thus in tuning to a source of correlation, the dendrites of a particular pyramidal cell find mathematical functions that relate that source of correlation, not only to sensory inputs, but also to other sources of correlation via intracortical connections. In this way, the connections among cells gain characteristics that dispose flows of activity to mirror regularities involving these sources.[8] This is extremely useful. When a SINBAD cell is activated, this amounts to the network "inferring" the presence of a particular source of correlation, both directly from sensory input, and indirectly from other cells that are active due to the presence of the source of correlation to which *they* have tuned. A cell that has tuned to a particular variable has a large number of sources from which it can obtain information about that variable, from numerous sensory input channels and also neighbouring cells. (This is due to the inductive richness of sources of correlation, and the capacity of a cell's dendriites to make use of much of this richness in learning to match.) If one of those sources of information is blocked, e.g. sensory inputs, the others will compensate.[9] In the context of the networks' dynamic isomorphism to the environment, a cell that corresponds to the kind *tiger* (because that is what it has tuned to) will light up when all that is seen is a twitching tail, or even a footprint. That is, intracortical connections allow the network to perform the trick of "filling in missing information".

Here then, in brief summary, is how SINBAD networks operate. The multiple dendrites on a SINBAD cell must find mathematical 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, richly endowed with internal links, comes to be dynamically isomorphic to the

---

[7] If there is no correlation available, the cell's activity will be low, and it will elaborate its dendrites in search of new inputs until it is able to find a correlation (for a review of dendritic growth, which occurs throughout life, see Quartz and Sejnowski, 1997). If it is still unsuccessful, at some point the cell will "give up" and degenerate (Edelman, 1987).

[8] 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. (There is also a mechanism to prevent too many cells from tuning to the *same* source of correlation – see Favorov and Ryder, forthcoming.)

[9] Of course, this will create a mismatch between dendrites; if a previously correlated input is *consistently* absent, the dendrite will learn to ignore it in order to achieve a match again with the other dendrites on the cell.

environment from which it receives inputs. This dynamic isomorphism mirrors the deep structure of the environment, with elements that correspond not only to sensory features, but also to the kinds (and other sources of correlation) around which environmental regularities are structured.

5. A teleoneurosemantics for the representation of kinds

Can the SINBAD theory explain how the representation of kinds is possible? It should be uncontroversial that the cortical network, if it is a SINBAD network, is a model building machine. Clearly the cortical network is supposed to structure itself isomorphically with regularities in the environment; the utility of this isomorphism is undeniable, for filling in missing information about the world, and in practical reasoning. But our main question, of how it is possible to represent kinds, turns upon the nature of the specific design principles of a SINBAD network. The design principles of a model making machine dictate its general function, and thus what type of structure it represents. We have seen that SINBAD networks have a strong *tendency* to become dynamic isomorphisms that mirror regularities organized around sources of correlation. The result we now want to get is that this tendency is teleofunctional: that the cortical SINBAD network was *designed* to develop isomorphisms so described, and consequently that SINBAD cells are *supposed* to correspond to sources of correlation. It would follow that they represent sources of correlation, and since kinds form one type of source of correlation, we would have shown how it is possible to represent kinds.

Since evolution is the designer here, we need to make it plausible that the SINBAD mechanism was selected for the properties of its interaction *specifically* with sources of correlation, that its being structured by sources of correlation in particular confers some benefit compared to other types of model-building (e.g. pairwise association). This is eminently plausible. We saw that 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. The result is an extremely powerful predictive network, with multipotent capabilities for filling-in.[10] Importantly, SINBAD cells must tune to *reliable sources of multiple correlations* in order for the network to exhibit this sort of power; the particular advantage of the network depends entirely upon the inductive richness of sources of correlation. SINBAD cells are plausibly built (by evolution) to take advantage of this inductive richness – they have a strong tendency to tune to sources of correlation, and this tendency is what ultimately produces a rich isomorphism.

[10] Note that several functions may overlap on a single dendrite, and typically cells will operate in *population* units, with all members of one population corresponding to the same source of correlation. So the capacity for filling in is astronomical (Ryder, forthcoming).

There are two aspects to the way in which SINBAD cells take advantage of the inductive richness of sources of correlation.[11] First, this richness permits a SINBAD network, given the nature of its units, to develop a correspondingly rich inductive network, not of scattered pairwise correlations (as in an associative network), but of interrelated regularities grounded in the deep structure of the environment. This richness ensures robust prediction through *redundancy* – there are many ways to predict the same thing. Second, the inductive richness of sources of correlation facilitates future learning. Because sources of correlation are inductively rich, once a SINBAD cell starts to tune to one by discovering *some* of the correlated properties it exhibits (the correlational seed), the cell (given its special properties) is in a uniquely advantageous position to discover *further* correlation. (This will be the case as long as its dendrites have not come to match their activities perfectly, which they almost never will, due to the presence of noise in the cortical network). We saw that in this way, a cell continually adds to its lines of "evidence" for the presence of the source of correlation to which it is tuning, indefinitely enriching the model's isomorphism.

So just as the automatic scale modeller has a dispositional "fit" for producing spatial isomorphs (and not density isomorphs), a SINBAD network has a dispositional "fit" for producing isomorphs to regularities centred around sources of correlation (and not to more generic sorts of regularities). It is precisely this fit with sources of correlation that gives SINBAD networks an advantage over other types of mechanisms with respect to richness and redundancy of prediction. So it is reasonable to infer that at least part of the cortical network, if it is a SINBAD network, was designed by evolution to come to mirror regularities specifically involving sources of correlation. This is one of the design principles of the cortical model making machine. In particular, SINBAD cells were designed to tune to and come to correspond to sources of correlation in the context of the SINBAD network's isomorphism. Since SINBAD cells have the general *function* of corresponding to sources of correlation, they *represent* or stand in for sources of correlation in the models the cortex produces. And since kinds are sources of correlation, SINBAD cells can represent kinds.

These general design principles, in conjunction with the history of production of a particular model, ought to allow us to determine exactly what that model represents.[12] In order to figure out which *specific* regularity structure is represented by a *specific* cortical model, we need to be able to identify what regularity structure served as a template for that model. Equivalently, we can figure out what sources of correlation served as the template for production of each of the *elements* of the model (i.e. particular cells that have tuned or are in the process of tuning to a source of correlation). The elements of a SINBAD model are particular cells that have tuned or are in the

[11] They correspond to two functions that Millikan attributes to empirical concepts (in her contribution to this volume, and her [2000] Ch. 3): "accumulating information" about kinds (and other substances), and "applying information previously gathered" about substances.
[12] Here follows a brief account. For a more detailed story, see Ryder (forthcoming).

process of tuning to a source 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. Something serves as a template for model production only relative to the design principles of the model-building machine.

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 due to correlations in some of the properties that have helped cause it to fire.[13] These properties are correlated due to their being properties of this particular source, so the cell finally comes to tune to and correspond to that source of correlation. This is the functionally normal route for a SINBAD model to adopt a particular configuration, the way it was designed to work: some specific source of correlation causes (or explains) each cell's achievement of dendritic matching. It is equivalent to the automatic scale modeler taking in an object though its input door, producing a nice mould, and spitting out a perfect model.

When an object causes structuring of the model by a functionally *abnormal* route, in a way that does not accord with the model-building machine's design principles, that object is not a template for the model. Thus when a rock falls in through the output door in the automatic scale modeler and causes a lump on a model of The Spirit of St. Louis, the result is not a model of the rock, but rather an imperfect model of The Spirit of St. Louis. In a SINBAD model, deviations from the way structuring is supposed to proceed by design will be deviations from the way tuning is supposed to proceed by design. These will include causal interactions with things that have inhibited a cell from achieving its current level of matching success.

For example, consider the following history of SINBAD model production. Suppose a cell had been gradually tuning to cats. Perhaps a dog caused the cell to fire at some point, because in certain conditions, dogs look like cats. Let us say that the dog made three of the cell's dendrites match, while there was a failure to match for two dendrites. The SINBAD learning rule made one of these dendrites modify its connections so that it became more responsive to a function that picks out dogs (or rather that dog in those circumstances), and *less* responsive to a function that picks out cats, a function that it had previously been tending towards. This *reduces* the cell's overall matching

[13] Despite the fact that the causal relation between the activity in a SINBAD cell's dendrite and a source of correlation is mediated by some intervening physiology, it is still causation in virtue of some determinate property of the stimulus. Which property was causally relevant to the activity in the synapse can be identified by counterfactuals. Suppose an instance of a particular shade of red causes a cell to fire. Had the stimulus been a different shade of red, the cell would have fired anyway. However, had the stimulus been blue, it would not have fired. Then the property that was causally relevant to the synaptic activity was redness, not the particular shade of red, nor colouredness.

success, but subsequent exposures to cats bring that dendrite back towards the function that picks out cats, and the cell back toward better matching success (and thus predictive utility). That response to a dog *inhibited* the cell from achieving its current level of matching success; it led it away from finding the correlated functions due to cats, the isomorphism it has now settled on. The dog was something that *affected* the model, but not according to design. It features in the history of the cell's tuning, but it does not cause or explain its matching success, i.e. the aspect of model structuring that occurs by design. It was a stray rock in the SINBAD mechanism, while the kind *cat* was this cell's template.

This does not mean that the model-building machine was *broken* when it changed so as to reduce its "predictive" utility; it was just functioning sub-optimally. Also note that our conclusion that this cell represents cats is consistent with an alternative history in which the cell permanently veers off its course, eventually tuning to and representing dogs. In this case, the cell would be in a different model, with a different history – and at some point in its progress, the kind *cat* may cease to explain its matching success. On the other hand, if the kind *cat* continues to explain its matching success, the cell will be disjunctive or "equivocal", where two kinds are confused as being the same (Ryder, forthcoming). (See Millikan [2000] on equivocal concepts, which certainly exist and so ought to be psychosemantically explicable.) This is another way model design can proceed sub-optimally. It is sub-optimal since it will lead to inductive errors.

So 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 its current 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.

So not only can SINBAD cells have the function of corresponding to kinds in the context of an isomorphism, the details of the SINBAD mechanism allow us to determine exactly *which* kind (or other source of correlation) a particular SINBAD cell has the function of corresponding to. If all goes well, that kind will be the unique representational content of the cell. Since SINBAD cells are the basic elements of a SINBAD network, we can also determine which regularity structure a whole network has the function of being isomorphic to, and thus models. Because of their inductive richness and SINBAD's penchant for such richness, kinds will tend to figure prominently in these internal models. Which, in addition to the evidence closely linking the cortex to the mind, is an important reason to suppose that mental representation, at least in us, is SINBAD representation.

References

Abeles, M. 1991: *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.

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

Aydede, M. 1997: Has Fodor really changed his mind on narrow content? *Mind and Language*, 12(3/4), 422-458.

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

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

Billman, D. O., and 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.

Block, N. 1986: Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy*, 10, 615-678.

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.

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. 1989: *Meaning and Mental Representation*. Cambridge, Mass.: MIT Press.

Cummins, R. 1996: *Representations, Targets, and Attitudes*. Cambridge, Mass.: MIT Press.

Douglas, R., & Martin, K. 1998: Neocortex. In The Synaptic Organization of the Brain. Shepherd, G. M. (Ed.), pp. 459-509. Oxford: Oxford University Press.

Dretske, F. 1981: *Knowledge and the Flow of Information*. Stanford: CSLI Publications.

Dretske, F. 1986: Misrepresentation. In Belief: Form, Content and Function. Bogdan, R. J. (Ed.), pp. 17-36. Oxford: Oxford University 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 Somesthesis and the Neurobiology of the Somatosensory Cortex. Franzen, O., Johansson, R., & Terenius, L. (Eds.), pp. 395-408. Basel: Birkhauser.

Favorov, O. V., and Ryder, D. forthcoming: SINBAD: a neocortical mechanism for discovering hidden environmental variables and their relations. *Biological Cybernetics*.

Fodor, J. 1981: The Present Status of the Innateness Controversy. In RePresentations: Philosophical essays on the foundations of cognitive science. 257-316. Cambridge, Mass.: MIT Press.

Fodor, J. 1983: *The Modularity of Mind*. Cambridge, Mass.: MIT Press.

Fodor, J. 1990: *A Theory of Content and other essays*. Cambridge, Mass.: MIT Press.

Fodor, J. 1991: Reply to Antony and Levine. Meaning in Mind: Fodor and his Critics. Loewer, B. (Ed.), pp. 255-257. Oxford: Blackwell.

Fodor, J. 1994: *The Elm and the Expert*. Cambridge, Mass.: MIT Press.

Fodor, J. 1998: *Concepts: where cognitive science went wrong*. Oxford: Oxford University Press.

Harman, G. 1987: (Nonsolopsistic) conceptual role semantics. In Semantics of Natural Language. Lepore, E. (Ed.), pp. 55-81. New York: Academic Press.

Jennings, D. L., Amabile, T. M., & Ross, L. 1982: Informal covariation assessment: data-based versus theory-based judgments. In Judgment under Uncertainty: Heuristics and Biases. Kahneman, D., Slovic, P., & Tversky, A. (Eds.), pp. 211-230. Cambridge: Cambridge University Press.

Keil, F. 1989: *Concepts, kinds, and cognitive development*. Cambridge, Mass.: MIT Press.

Kornblith, H. 1993: *Inductive Inference and Its Natural Ground*. Cambridge, Mass.: MIT Press.

Matthen, M. 1988: Biological Functions and Perceptual Content. *Journal of Philosophy*, 85, 5-27.
McAllister, A. K. 2000: Cellular and molecular mechanisms of dendrite growth. *Cerebral Cortex*, 10, 963-973.
McGinn, C. 1989: *Mental Content*. Oxford: Blackwell.
McLaughlin, B. 1987: What is wrong with correlational psychosemantics. *Synthese*, 70, 271-286.
Medin, D. L., & Ortony, A. 1989: Psychological essentialism. In Similarity and Analogical Reasoning. Vosinadou, S., & Ortony, A.Cambridge: Cambridge University Press.
Millikan, R. 1984: *Language, Thought, and Other Biological Categories*. Cambridge, Mass.: MIT Press.
Millikan, R. 1998: A common structure for concepts of individuals, stuffs, and real kinds: more Mama, more milk, and more mouse. *BBS*, 21(1), 55-65.
Millikan, R. 1999: Historical kinds and the "special sciences". *Philosophical Studies*, 95(1&2), 45-65.
Millikan, R. 2000: *On Clear and Confused Ideas*. Cambridge: Cambridge University Press.
Pakkenberg, B., and 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.
Papineau, D. 1993: *Philosophical Naturalism*. Oxford: Blackwell.
Phillips, W. A., and Singer, W. 1997: In search of common foundations for cortical computation. *Behavioral and Brain Sciences*, 20, 657-722.
Prinz, J. 2002: *Furnishing the Mind: Concepts and their perceptual basis*. Cambridge, Mass.: MIT Press.
Putnam, H. 1975: The meaning of meaning. In Language, Mind and Knowledge. Gunderson, K. (Ed.) Minneapolis: University of Minnesota Press.
Quartz, S. R., and 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 Dissertation, Chapel Hill: University of North Carolina, Department of Philosophy.
Ryder, D. forthcoming: SINBAD Neurosemantics: A theory of mental representation. *Mind & Language*.
Ryder, D., and 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., and Tanne, D. 1994: Perceptual learning: learning to see. *Current Opinion in Neurobiology*, 4, 195-199.
Stalnaker, R. 1984: *Inquiry*. Cambridge, Mass.: MIT Press.
Stampe, D. 1977: Toward a causal theory of linguistic representation. In II. French, P., Uehling, T., & Wettstein, H. (Eds.), pp. 42-63. Morris, MN: University of Minnesota.
Swoyer, C. 1991: Structural representation and surrogative reasoning. *Synthese*, 87, 449-508.
Usher, M. 2001: A statistical referential theory of content: using information theory to account for misrepresentation. *Mind & Language*, 16(3), 311-334.
Woolley, C. S. 1999: Structural plasticity of dendrites. In Dendrites. Stuart, G., Spruston, N., & Häusser, M. (Eds.), pp. 339-364. Oxford: Oxford University Press.