

they are at (even relative to one another) may or may not be represented. The same is true of the location of objects on the proximal stimulus (e.g., on the retina) or further up in the nervous system, such as patterns of activity on the retinotopically organized fibers leading from the eye, or in the primary visual cortex, which is largely retinotopically mapped. Since these locations are past the sensors, are they necessarily representations? If so, what is the essential difference between the way that distance in the world affects perception and the way that the corresponding distance on a neighborhood-preserving (i.e., homeomorphic) anatomical mapping affects perception (for ease of reference I will refer to the results of such mappings as “neural layouts” or NLs)? We can say that such neural layouts *register* (rather than represent) spatial properties. They help to illustrate the general theme that there are many types of representations, ranging from conceptual, through subpersonal, to informational states that are better referred to as registrations rather than representations. In the next subsection I will examine neural layouts to see if they warrant the conclusion that spatial properties are always represented in NLs since locations appear to be roughly preserved on a maplike surface.

(3) Are neural layouts always representations? Intuitively it seems that a neural layout (a layout of activity in the cortex that is a homeomorphic mapping of some other spatial domain, such as shown in figure 4.4) carries information about location in a special way that makes it a maplike representation (I will have more to say about maplike representations and their role in navigation in section 5.4.2). The intuition is that any projection of spatial information onto a neural layout (NL) is automatically a representation since it preserves spatial locations (at least to a first approximation). This intuition derives from the fact that such an NL resembles a canonical map or picture and could (if spread out) be used by a person to navigate or to recognize patterns. However, the layout need not be used in this way.

Whether we call any retinal or other neural layout a *representation* is partly a question of terminology, and NLs usually do carry information about something in the world to which they are causally connected. As mentioned above I prefer to call that type of mapping a *registration* of information—spatial properties are registered rather than represented in NLs. What does matter is not the terminology, but the distinctions we need to make with respect to the role NLs play in explanations. If we use the term “representation” to refer to any information-bearing state, then we will still need to distinguish another, stronger sense of representation. The main distinction we still need is that between states whose representational *con-*

tent plays a role in explanations and those in which the content (if any) does not play any role. If we gain no explanatory advantage by specifying *what* an NL represents, then nothing is gained by treating the NL as a representation. The fact that the NL looks like a map—even if places on the NL correspond to places in the world—is still not enough for it to be a representation in the strong sense.

There are several specific requirements that should be met for something to count as a spatial representation in the strong sense. We need to show not only that locations, distances, and directions in the NL correspond to the same properties in the world but also that they determine the organism's behavior vis-à-vis those represented places. In other words, we need to show that these properties of the NL function to represent properties of the world for the organism. One indicator that they function in this way is if at least some generalizations concerning behavior require appeal to the represented domain as opposed to the pattern of the NL itself. Some principles governing NLs might well be captured solely in terms of properties of the NL with no regard to what they may represent. The principles for forming clusters of features and most Gestalt grouping principles may well be of this sort. These principles (at least as understood by people like Kohler and Wertheimer) are expressed over properties of the proximal stimulus or over neural fields in the brain,⁵ but not over locations, distances, and directions in the world.

One way to see this is to reflect on the fact that unless the function of the NL is to represent spatial properties for the organism, it would not be possible for the NL to *misrepresent* something. The possibility of misrepresenting is a signature property of representations—a retinal pattern cannot misrepresent the visual world since optics does not make “mistakes.” Similarly, it is meaningless to ask of an NL in which frame of reference it represents an

5. Processes operating over NLs typically respond to spatially local properties of the NL—they operate over “local support.” The principles of operation of such processes are *prima facie* expressible over nonconceptual neural properties. Recently there has been an increase of interest in applying dynamic systems theory to modeling the mind. Since such theories are generally not representational (and not computational in the sense discussed in Pylyshyn 1984) there is little chance that they will explain cognitive processes. But they may find application in the sort of nonrepresentational processes that transform NLs or registrations, derive Gestalt clusters, solve the correspondence problem in certain cases, and even carry out tracking (examples are found in Koch and Ullman 1985; Pylyshyn 2003, appendix 5A). Thus theories that postulate spatial registrations may be appropriate for the sort of neural field processes envisioned by Wolfgang Kohler (1947).

object's location, since by itself it does not represent an object as located anywhere. But in the strong sense of representation, where the NL functions to direct movements or to identify objects, it *does* matter how its spatial relations are represented. In that case an NL may represent some locations with respect to a head-centered frame of reference, or as being to the left of another location, or as being more than an arm's length away; and for purposes of determining actions *it matters how the location is represented* (or what it is represented *as*). Without this strong sense of representation, with only a direct object-to-NL mapping, there is no possibility of misrepresentation, and thus it is misleading to call the NL a representation or a map.⁶

It's important to keep in mind that this discussion is about *explaining* regularities in vision and behavior. So the answer to the question at hand—whether an NL is a spatial representation—is that it depends on whether one must refer to the geometrical properties of the represented world in providing explanations. For example, do the principles (such as principles of clustering or of correspondence) that have to be cited refer to properties of the NL or properties in the world? Suppose, for the sake of argument, that the clustering algorithm applies only over distances on the neural lay-

6. I am leaving out a lot here. What makes a terrestrial map able to misrepresent is that this sort of map typically is constructed with the intention that certain of its features correspond to certain features of the relevant terrain, and the map has to be interpreted with these intentions in mind. Thus there is ample room for the intended correspondence to fail and for the map thus to misrepresent. These degrees of freedom are absent in the case of NLs unless we assume that the map is interpreted by some process that allows a possibility of misinterpretation. Sometimes it is tempting to assume an interpreter, and at other times it is tempting to assume a design purpose for the NL—and sometimes it makes sense to talk of a “map” even though there may be no NL, as in the case of insect navigation (see section 5.4.2). Talk about the design purpose (what the NL is *for*) is sometimes helpful, even though strictly speaking there is no agent who designed the representation-using system, because it ties together a variety of otherwise unconnected properties of various mechanisms. In fact our understanding of “natural constraints” rests on assumptions about the purpose of some of the mechanisms, and Marr (1982) motivated his analysis by asking what various visual mechanisms were *for*. Dretske (1981) suggests another way in which an information-carrying state might misrepresent, a way based on learning: If the system has been exposed to pairs of properties and internal states, it could learn which features of the environment the states represent and thus could be in a position to misrepresent those features. These are all questions that I will not get into, beyond arguing that there is more to being a map than homeomorphism.

out, which, in turn, is a homeomorphic transformation of activity on the retina.⁷ In that case nothing is gained by saying that these distances represent properties in the world, since by hypothesis the distance on the NL is all that is relevant to explanations involving distances and those are indifferent between whether it represents a visual angle, a 2-D distance on the retina, a 2-D distance far away from the observer, or a distance in 3-D oriented at the appropriate angle from the viewer to project onto the line on the NL. Therefore, it is not a representation in the strong sense; it does not represent the property *as* something in the world, notwithstanding that, if spread out on a flat surface, the pattern of activity looks like a map. But since it carries some information about spatial locations we say that it *registers* spatial properties.

(4) *When should we postulate representations?* The purpose of postulating representations is to provide explanations and to capture generalizations that would not be captured without reference to the contents of such representations—to what they represent. But sometimes (as in the hypothetical NL discussed above) the function of information-carrying states can be fully described without reference to contents. It could be that principles such as, say, those involved in clustering or apparent motion can be fully explained without reference to any representational content of the states involved. In discussing the way information might be carried by an NL, I noted earlier that an explanation might sometimes be stated in terms

7. These examples are for purposes of illustration; I am not prejudging the empirical question of whether the principles of clustering or of pairing features to solve the correspondence problem apply only to proximity on the NL. If they apply to distal properties then the present argument would not work—but then again neither could we claim that the NL is the basis for the clustering or correspondence solution, since we know at least that V1 (or any other NL) is prior to processes that provide 3-D information (prior to the constancies). There have been conflicting claims over whether 3-D properties are relevant to apparent motion; some investigators maintain that 3-D distance is relevant (Attneave and Block 1973; Wright, Dawson, and Pylyshyn 1987) and some that it is not (Ullman 1979). Recent years have seen many reports of 3-D properties being relevant to what seem like early processes, such as popouts in search (Enns and Rensink 1990; Rensink and Enns 1995) or even multiple object tracking, where it seems that speed in the distal world, rather than on the retina, determines the performance in MOT (Enns and Franconeri 2006; Liu, Austen, Booth, Fisher, et al. 2005). These suggest that such processes are postconstancy or postdepth analysis and therefore do not involve (only) places on the NL (in V1). But this is an empirical question that requires further research.

Various sorts of hybrids and combinations are possible. Maps, of course, typically have symbolic elements in addition to their strong iconicity. They can also have weakly iconic and quasi-iconic elements, as when using color to represent air pressure or wind speed. Weak iconicity requires vehicle isomorphism, which requires internal vehicle structure; so when colors quasi-iconically represent wind speeds, then *patterns* of color will weakly iconically represent *patterns* of wind speeds. The relation between particular colors and particular wind speeds is scheme-level and a matter of co-similarity. Iconic representations can also combine discrete and continuous elements. A budgetary pie chart, for example, usually represents budget categories discretely and symbolically and represents percentages continuously and (weakly) iconically.

The diagram claims that all continuous representations are at least quasi-iconic.¹⁴ I don't have a decisive argument for this, but I think it's probably true. All the instances I can think of are pretty clearly so. Additionally, I think that the noise inherent in continuous representations—the ease of mistaking one representation for a similar one—would be devastating if similar representations weren't constrained to have similar contents; thus, as a practical matter at least, continuity requires quasi-iconicity.

Again, this isn't about the *word* "iconic," and I wouldn't mind much if someone wanted to use it only for something much narrower, perhaps just those strongly iconic representations that represent in virtue of their spatial properties. What really matters are the distinctions illustrated by the Euler diagram. There are important differences between systems that represent by sharing properties and those that represent by having analogous properties. There are important differences between systems whose individual vehicles are similar to their representanda, and systems where the similarities only occur at the system level. And yet there are important similarities among all of these; co-similarity distinguishes all of these from symbolic representation and makes sense of the intuitive claim that iconic representation is somehow less "arbitrary" than symbolic representation.

4 Iconic *mental* representations

So far, I've deliberately focused on tangible, physical, non-mental representations, on paper maps and photographs and the like. My hope, however, is to connect this up with mental representation in an effort to better understand the latter. The most interesting question in this neighborhood, I think, is whether the mind employs any of the essentially spatial strongly iconic representations discussed above. I'll focus on maps, but I think the same considerations all apply to pictures as well. The idea that there are literally maps in the head now seems problematic, in a way that the idea that there are literally sentences in the head is not. Sentences are individuated only by their semantic and syntactic properties, so the thought that minds or brains could literally have sentences in them doesn't pose any fundamental trouble. It's just

¹⁴ I.e., everything that's "analog" in the sense I would prefer (i.e., not digital) is "analog" in the sense some other authors prefer (i.e., not symbolic). Though not vice versa.

the thought that mental or brain states (or events, etc.) could have the same syntactic and semantic properties as some sentences. But strongly iconic representations represent by sharing properties with their objects, which requires them to *have* the relevant properties—in this case, spatial properties.

Surely there's no obstacle to the mind or brain exhibiting the isomorphisms required for weak and quasi-iconicity. Maybe that's iconic enough, without having to worry about strong iconicity? Can't we just say that some mental representations are at least quasi-iconic, and leave it there? I want to defend a fairly articulated view about the relation between iconicity and perception: although I think it is implausible that all perception is iconic in even the weakest sense (quasi-iconicity), I think it is nevertheless plausible that some perceptual representations may be iconic in the strongest sense. It is thus worthwhile to explore the possibility of strongly iconic mental (perceptual) representation, rather than just dropping the issue and settling for quasi-iconicity.

There's nothing problematic about the idea that mental representations might have temporal properties, which it might share with representanda and thus represent temporal properties in the environment in a strongly iconic way. But I think an even stronger, less obvious claim is defensible; I think the idea of strongly iconic *spatial* representations in the brain is not hopeless. Making sense of spatial vehicles in the brain is a difficult project that I can only touch on here, but it's a worthwhile project, as I hope to show below.

This difficult project, however, should be distinguished from a number of easy solutions that present themselves, all of which, I think, are dead ends.

4.1 Some dead ends

We want to try to make sense of the idea that mental representations could literally have spatial properties. One obvious possibility is to invoke sense-data. Sense-data are hypothesized mental items that literally have the very properties that external objects perceptually appear to have. For the table to look white, or rectangular, is for it to present a sense-datum that really is white, or rectangular. I won't say much about sense-data here, except to say that anyone who believes in them has a straightforward route to strong iconicity.

Another easy way out of the problem would be to simply appeal to phenomenal spatial properties. Suppose there's no special problem for the idea of phenomenal rectangularity and the like. Then maybe for any spatial property *F*, a mental representation “with” *F* (as opposed to merely *of F*) could just be a state that's phenomenally *F*? One problem with this move is that the question of whether (any) perceptual representations are iconic was supposed to be an empirical, scientific question, not one for introspection. The question wasn't supposed to be about how things seem to us, but about why they seem that way. Furthermore, a good many of the mental representations thought to perhaps be iconic are not (phenomenally) conscious, so phenomenal spatiality wouldn't help us with these questions anyway.

There is a lot of talk among neuroscientists about “cognitive maps” (Tolman, 1948; O'keefe & Nadel, 1978; Derdikman & Moser, 2011), especially in the

hippocampus and related structures. What's meant by virtually all such talk, however, is that certain brain states encode navigation-enabling information about what is where. These internal representations have the *content* of maps and are "map-like" in the important sense that they carry information about metric relations among items in the world, and not just which direction home is. But this doesn't speak to our question here, which is about the *format* of these representations. Just as a subway rider could use a verbal list of the stations in order, rather than an actual *map* with this information, the mere fact that, say, place cells in the hippocampus carry information about conjunctions of features found in particular environmental locations, doesn't begin to argue that this information is encoded in a spatial, map-like format. But the format matters, as the subway map example illustrates: the very same information (content) is much easier to use when encoded in a map format, rather than a verbal one.

One last easy way out, which seems to have tempted nearly everyone who's written on the subject, is to note that the brain employs a large number of topographic maps. There are retinotopic maps throughout the early visual cortex, where adjacent brain regions code for adjacent retinal areas. There are tonotopic maps in the cochlea and auditory cortical regions, where high pitches are coded at one end and low pitches at the other, with in between pitches in between. And so on. These are highly significant phenomena, of course, but they don't have anything to do with the iconicity of mental representations. Of course the brain has spatial properties and some of these might be shared with distal objects; the question is whether *mental representations* might have spatial properties, and topographic maps do nothing to answer this question.

Take retinotopic maps. In a seminal study (Tootell et al., 1988), a monkey was given a radioactive dye while looking at a bullseye pattern; afterwards, the monkey's primary visual cortex was placed on a radiation-sensitive photographic film, where a (distorted) bullseye pattern was clearly visible. There were literally lines, curves, and wedges in the monkey's brain. (Or so let us concede.) Yet, there's no reason to think that at this very early stage of processing—V1—anything is being represented *as wedge-shaped*. In fact, there are reasons to think not, i.e., to think that lines, curves, wedges and the like are represented *as* lines, curves, etc. in different brain regions and later. The topographic map is often said to be a "representation" of the activity of retinal photoreceptors. Yet there's no reason to think that the brain *ever* forms a single unified mental representation of the retina. It's not needed, so long as a large number of highly coordinated representations of very local states of retinal activity exist. The contents at this level aren't things like "curve," "square," "wedge," but rather "edge at retinal location x ," etc. Their topographic organization presumably facilitates the coordination of these local representations (as well as contrast enhancement by way of lateral inhibition, etc.), but the representations are still local.

Second, it is well known that, due to cortical magnification, the topographic maps are distorted: there are more cortical neurons with foveal receptive fields than with peripheral receptive fields. Consequently, a square presented to one side of the fixation point will produce a trapezoid in the cortex, the side nearer to the fixation point receiving a larger share of cortex. But if strongly iconic representations represent

by having the very properties they attribute, this raises a puzzle: why shouldn't this represent the object as being trapezoidal, or projecting trapezoidally? There's nothing physically rectilinear in the cortex, so no strongly iconic representation as of rectilinearity.

The solution, of course, is what we knew anyway, that it's functional spatiality, rather than physical spatiality that's doing the representational work. If we could rearrange and jumble these cortical neurons while keeping their connectivity and timing the same, the topographic organization would disappear, but all the semantic and syntactic properties of the mental representations would stay the same, including the iconicity or not of these representations. Topographical realization is simply irrelevant to the question; we have to seek iconicity at the representational level, not the implementational level. All this means we can't avoid the task of coming to a better understanding of what functional spatiality—which I equate with spatiality at the level of the representation, rather than the realization—might mean.

4.2 Functional implementation

Let a *representational scheme* consist of (a) a set of representations and (b) a semantics for these representations. In compositional schemes, (a) is specified by listing primitive representational elements and providing a set of combination rules for generating well-formed representations out of these, and (b) is given by specifying contents for the primitives along with semantic composition rules to generate contents for the complex representations. Schemes will be individuated rather finely. Two schemes could differ even while having the very same “syntactic” elements (the same primitives and same possible combinations among them), if they had different semantics.

Much of cognitive psychology is concerned with discovering what kind of representational scheme a given system is using/implementing. This research, I think, is guided by the following considerations. To implement a representational scheme of a given type (e.g., Venn diagram, subway map, Roman numerals, etc.) the implementation needs to match the scheme in at least the following, functional, ways:

- (i) It needs to have the same *expressive power* as the scheme. Standard Venn diagrams, for example, can express relations among three predicates, but no more; any system that can express more is not implementing a (standard) Venn diagram scheme. If pictures do have distinctive skeletal contents, then a system can't be implementing a pictorial scheme if its states don't have those contents.
- (ii) It needs to exhibit the same *systematicities*. As Fodor and Pylyshyn (1988) famously noted, cognitive capacities tend to come in clusters. Among other things, this means that as certain representational capacities are gained or lost, they do so in groups. Gaining a new primitive automatically brings with it a host of new complexes; losing a combination rule means losing all the complex representations that rule made possible, etc. As Cummins (1996b) points out,

neural activity to occur. We can take Maley's summary of Zacks' review findings at face value. The problem is that we still do not know what it is about these representations' use of magnitudes to represent magnitudes that predicts the response times.

5. Deep Neural Networks: An Alternative View of Retinotopic Organization

Even if the phenomenon of retinotopic organization cannot be connected with the mental rotation and mental scanning timing results in a way that supports the thesis that mental images are analog, some may see the observation of retinotopic cortical activation as sufficient evidence by itself for the thesis the mental images are analog (as in, e.g., Nanay [2023], p. 48)). After all, why would the brain bother to generate image-like patterns of activation while representing patterns in the world if it were not in fact making use of imagistic representations?

However, there are many reasons the brain might generate image-like patterns of activation even if it were not making use of analog, imagistic representations. An especially compelling reason has been obscured by the fact that the imagery debate's participants often assume a false choice: that mental imagery must either be *iconic* or *language-like* in its representational format. These, for many of the key participants, were the only formats worth serious consideration. That is surprising, given that connectionist alternatives—which invoke neither iconic nor language-like representations—were well-known and already much-discussed in the 1980's and 1990's, when the imagery debate was in full swing. The explanation, I suspect, is in part technological and in part sociological. At the time, there were no connectionist AI systems remotely capable of the kinds of image generation tasks we associate with imagery; while, simultaneously, the most influential researchers on both sides of the debate had in common a dim view of connectionism. Pylyshyn, in particular, is famous for his independent attack on connectionism as a theory of thought (Fodor & Pylyshyn [1988]).

Things are very different now. Today's sophisticated image classification and image generation AI systems all rely upon deep neural network (DNN) connectionist architectures. There are, by contrast, no AI systems of remotely comparable capacities that make use of iconic or language-like representations (*modulo* the assumption—which cannot be fully defended here—that the DNNs in question do not implement iconic or language-like representations). Thus, the possibility that DNN-like connectionist architectures underlie human vision and mental imagery deserves very close consideration.

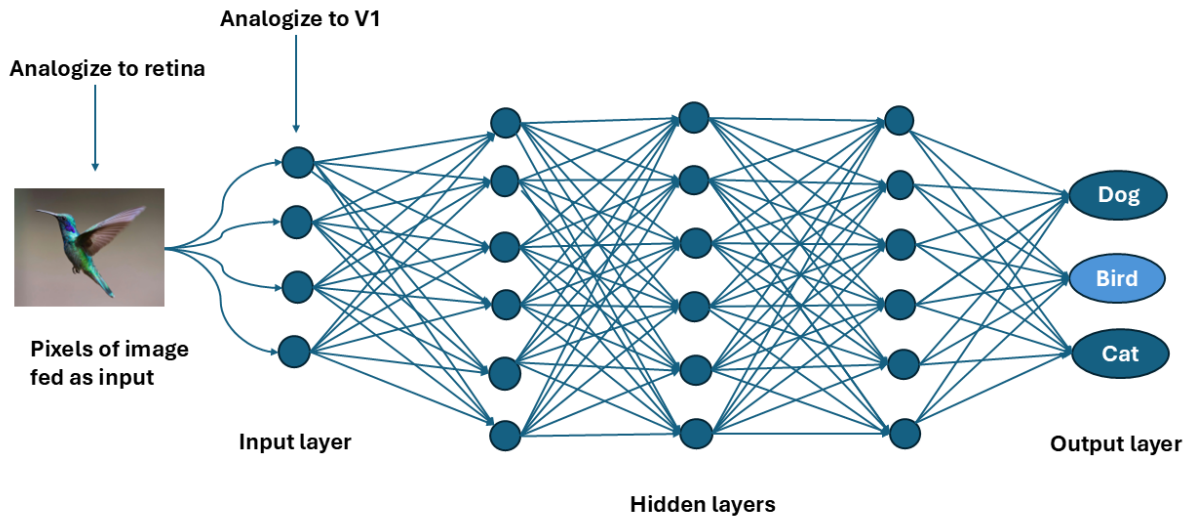


Figure 2: A representation of a convolutional deep neural network showing a characteristic lack of connection among nodes within any single layer.

A central selling point of connectionist architectures, considered as models of human cognition, is that the connections among layers of nodes in an artificial neural network bear clear structural similarities to the connections among neurons in the human brain. Accordingly, the important feature of image-processing DNNs for our purposes is that they offer an account of why we might see retinotopic activation in the brain during imagery (and visual perception) tasks, even if such activation did not constitute the tokening of iconic mental representations. Consider the feed-forward DNN shown in Figure 2, where an image of a hummingbird activates an input layer of four nodes, whose activations feed forward to nodes in three hidden layers, resulting in the activation of three nodes in the output layer. When the central node in the output layer is activated to a sufficient threshold, the network counts as indicating that a bird was in the input image. We can analogize this processing to human neural activation by letting the hummingbird image—considered as a 2D grid of pixel activations—stand for activation at the retina, and the input layer (also structured as a 2D arrangement of nodes) analogized to early visual areas that show retinotopic activation. If these analogies hold even roughly, it would be unsurprising to see structurally corresponding activations at the retina and at the input (and other early) cortical layers, simply because cells in the retina are activating subsequent neurons in early visual areas in ways corresponding to their own spatial outlay. Importantly, however,

when we conceive of what such activations are accomplishing computationally in connectionist terms, we see that there are no iconic (or depictive, or even analog) representations at work.

There are several interrelated reasons for this. First, it is only thanks to the cascade of activation across multiple layers—eventuating in the output layer’s middle node activating above a threshold—that (say) a *bird* is represented. The cognitive system cannot be said to represent—by distinguishing from other stimuli—the sort of thing causing activation at the input layer until the nodes at other layers have played their role in shaping activation at the output layer. Further, the processing in connectionist networks is *parallel* in the specific sense that each node within a layer is processing inputs from each node in the prior level and sending a signal forward to each node in the subsequent layer *completely independently* of what else is going on with other nodes in the same layer. Critically, *there are no connections between nodes in the same layer* (as shown in Fig. 1). Because there are no connections among nodes within a particular layer, it makes no difference how far apart those nodes are from each other. It would not change the network’s functioning at all to radically alter the space between nodes in a layer—making it entirely non-uniform, with any apparent retinotopic image disappearing—so long as each node maintained its connections to prior and subsequent nodes. In contrast, if spatial magnitudes were really being used to represent spatial magnitudes—if there really were iconic representations present within the networks—such changes should make *all the difference* to what is being represented.

In sum, the striking appearance of image-like activation in retinotopically organized cortex gives us no reason to suppose that the brain is making use of iconic representations. A brain whose processing mirrored the principles at work in our most sophisticated image processing and image generating artificial intelligence would also show such activations, without their being iconic representations. This is just one salient example of likely many for why we might see retinotopic activations in a system that is not using iconic (or analog) representations.

6. A Format-Agnostic Account of the Timing Results

What, then, explains the response times we see in the rotation studies? I end with a proposal that, while schematic, suggests that questions of format will not be at the fore. While I