Journal of Experimental Psychology:
Human Perception and Performance
1995, Vol. 21, No. 2, 423–431

On Computational Evidence for Different Types of Spatial Relations Encoding: Reply to Cook et al. (1995)

Stephen M. Kosslyn and Christopher F. Chabris
Harvard University

Chad J. Marsolek
University of Arizona

Robert A. Jacobs
University of Rochester

Olivier Koenig
University of Lyon 2

Computational models in psychology play an increasingly important role in characterizing theoretical distinctions, understanding empirical results, and formulating new predictions. However, the proper use of models is subject to debate and interpretation, as Cook, Früh, and Landis (1995) have demonstrated in a critique of neural network simulations reported by Kosslyn, Chabris, Marsolek, and Koenig (1992). These simulation results supported a distinction between two types of spatial relations encoding. Cook et al. argue that Kosslyn et al.'s models did not process "spatial" representations and that input–output correlations rather than properties of spatial relations encoding processes explain the performance of the models. This article provides conceptual and analytic rebuttals of those criticisms.

Humans rely on vision for numerous purposes, some of which make contradictory demands on our information-processing systems. For example, to reach for objects and navigate among them, one needs to know their precise distances; to recognize and identify objects, however, it is better to ignore this information. In contrast, abstract spatial relations (such as "connected to") among the parts of an object can be important for recognition and identification but not for reaching and navigation. Such considerations led Kosslyn, Chabris, Marsolek, and Koenig (1992) to propose that the brain represents spatial information in more than one way, with different types of spatial representations being useful in different circumstances (see also Jacobs & Kosslyn, 1994; Kosslyn, 1987). Specifically, we have argued that *categorical spatial relations,* such as above–below, left–right, or on–off, group a set of relative positions into an equivalence class. These representations are useful during object recognition and identification because they can specify invariant relations among parts (e.g., the fact that the forearm remains "connected to" the upper arm even when the arm is bent in different positions). In contrast, *coordinate spatial relations* specify precise metric distance; these representations are primarily useful for guiding motor movements. The information that is preserved in coordinate spatial relations representations is discarded in categorical spatial relations representations.

An appreciation of the complexity of the brain and its role in human behavior has led many researchers to two fundamental insights: First, it is unlikely that a natural language, such as English, will allow us to express detailed theories of brain function. Rather, the conception of the brain as a computing machine has led researchers to adopt the vocabulary of computation to describe brain function (e.g., Churchland & Sejnowski, 1992; Kosslyn & Koenig, 1992). Second, informal descriptions of theories about the brain can often explain both a result and its converse. (For an example of this point applied to the study of cerebral lateralization in visual processing, see the comments of Marshall, 1981, on Nebes, 1978.) Thus, a primary virtue of computational theories is that they can be instantiated in precise models of cognitive function (implemented as computer programs), which can be tested directly. We have used results from computational models as one source of evidence for the distinction between two types of spatial relations encoding, and our interpretation of these results has recently come under fire from Cook, Früh, and Landis (1995).

As Cook et al. (1995) have made clear, the use of computational models is not always as simple or straightforward as it might appear. Not only must many decisions not relevant to theory be made in order to build a working model (see Hesse, 1963; Kosslyn, 1980) but some features of a model often can be interpreted in several different ways. Because of the inherent ambiguities in interpreting models, we have consistently adopted a converging evidence approach to studying the nature of human spatial relations encoding; we have examined not only the conver-

Stephen M. Kosslyn and Christopher F. Chabris, Department of Psychology, Harvard University; Chad J. Marsolek, Department of Psychology, University of Arizona; Robert A. Jacobs, Department of Psychology, University of Rochester; Olivier Koenig, Department of Psychology, University of Lyon 2, Lyon, France.

Preparation of this article was supported by Grant 2 PO1 17778-09 from the National Institutes of Health, Grant N00014-94-1-0180 from the Office of Naval Research, and by a National Science Foundation Graduate Fellowship. We thank David Baker, Joe Cruz, and Joel Pust for their assistance.

Correspondence concerning this article should be addressed to Stephen M. Kosslyn, Department of Psychology, Harvard University, Cambridge, Massachusetts 02138.

gence in results from different types of models but also the relation between results from models and data obtained from experiments with human subjects.[1] Although there is ample empirical evidence that categorical and coordinate spatial relations are encoded by at least two different processes, these findings do not speak directly to the issues raised by Cook et al. They challenged our assertion that results from our computational models can be treated as an independently valid source of evidence, and we answer that challenge in this article.

Cook et al. (1995) argued that the conclusions Kosslyn et al. (1992) drew from their neural network simulations are flawed in two fundamental ways. First, Kosslyn et al. conceptualized the tasks given to network models as (categorical and coordinate) spatial tasks, but Cook et al. argued that the networks do not actually process spatial information. Second, Cook et al. argued that the performance of the networks can be completely understood by the peculiar correlational structure of their inputs and outputs, which has nothing to do with a possible distinction between categorical and coordinate spatial relations encoding. They argued that the performance of the Kosslyn et al. networks simply reflects the availability in the input stimuli of "definitive" information, or the degree to which activation of some input units is consistently associated with only one response.[2] We address the two fundamental points made by Cook et al. (1995) and conclude with some general observations about the role of computational modeling in a larger program of empirical research.

Spatial Properties in Neural Network Models

Cook et al. (1995) placed much stock in their claim that the Kosslyn et al. (1992) network models did not process spatial information. They showed that networks with scrambled input arrays produced similar results to those that do not have scrambled input arrays, and they concluded that this result shows that the input array is not spatially organized. As discussed below, Cook et al.'s critique rests on infelicitous assumptions about the nature of spatial representations and the structure of receptive fields.

Spatial Representations

Consider the fact that a photoreceptor located above another photoreceptor in the retina actually encodes a location perceived as below that of the latter cell. One way to demonstrate this fact is to stimulate the receptors and determine whether the subject sees one location as above or below the other (i.e., examine how input vectors are mapped to spatial judgments on the part of the human subject). To establish this mapping, we need to know the spatial judgments (the output) and how they are correlated with stimulation of retinal cells (the input). Only by examining this relation can we determine anything about the functionally spatial properties of the information in the array of retinal cells.[3]

[1] The empirical evidence is derived from three methods, the first of which is divided visual field studies. For example, Hellige and Michimata (1989) and Kosslyn, Koenig, et al. (1989) lateralized stimuli that contained a bar and a dot, and asked the subjects to decide whether the dot was above or below the bar or whether it was within a criterion distance of the bar. Subjects made the metric judgment faster when the stimuli were presented in the left visual field (and hence were encoded initially by the right hemisphere) and tended to make the categorical judgment faster when the stimuli were presented in the right visual field (and hence were encoded initially by the left hemisphere). Indeed, Kosslyn, Koenig, et al. (1989) showed that subjects judge whether objects are within 1 in., .5 in., or 1 cm faster in the right hemisphere, and evaluate on–off, above–below, or left–right faster in the left hemisphere. (These basic findings have been obtained many more times, e.g. Hellige et al., 1994; for a review, see Kosslyn, 1994; cf. Laeng & Peters, in press). The dissociation between hemisphere and judgment would not occur if only one type of process were used to encode both types of spatial information. The second method focuses on deficits following focal brain lesions. Researchers have documented selective deficits for the two kinds of spatial relations encoding following damage to the left or right cerebral hemispheres (e.g., see Laeng, 1994). The third method is brain imaging. Our group has recently used positron emission tomography to show that the posterior left hemisphere is more active when subjects encode categorical spatial relations than when they encode metric spatial relations, but vice versa for the posterior right hemisphere (Kosslyn, Gitelman, Thompson, Rauch, & Alpert, 1994).

[2] Cook et al.'s (1995) critique also rested partly on the notion that the hidden-layer units in the models failed to abstract general principles of spatial processing to solve the problem of encoding spatial relations. Cook et al. asserted, regarding an example of their own devising but meant to be analogous to Kosslyn et al.'s (1992) networks, that "The net now shows little ability to generalize because the hidden units do not encode *general concepts* [italics added] of the geometrical structure of the input patterns" (p. 412). This conclusion would appear to be supported by the results of Scalettar and Zee (1988), who found that a three-layer feedforward network did not learn a single "generalizable" representation of a categorical spatial relation (left–right), but instead developed representations of specific exemplars. However, although the Scalettar and Zee results are intriguing, we should point out that their network architecture used as many hidden units as input units (10 each; see Scalettar & Zee, 1988; p. 193), whereas that of Kosslyn et al. (Study 1, Part 1; p. 566) used between 4 and 10 hidden units in the categorical partition with a 28-unit input array. Furthermore, the Scalettar and Zee training procedure involved drilling the network on input–output pairs on which performance was poor (Scalettar & Zee, 1988, p. 194). Both of these practices (large hidden layers and drilling) may foster the development of exemplar-based representations at the hidden layer and were not followed by Kosslyn et al. Thus, the conclusions of Scalettar and Zee do not necessarily apply to the Kosslyn et al. studies.

In addition, Kosslyn (1980, 1994) reviewed evidence that people can shift the focus of attention, even if no eye movements are made. By shifting the location of attention, one can essentially translate the input region of the network to different portions of the visual field, which will help the network to encode spatial relations in a wide range of positions, even if no general concept is represented.

[3] Cook et al. (1995) wrote "The essential question is whether the Kosslyn et al. neural nets, when making judgments about such inputs, actually detect spatial information in the input stimuli or perhaps detect only specific combinations of on–off units, and

The same is also true in network models. To understand the function that a network computes, we must determine how input vectors are correlated with output vectors. For the original 80 unscrambled vectors in Cook et al.'s (1995) models (see Cook et al., Table 3), the coefficients are all 1.0 for correlations between Input Units 1–9 and the "above" o.ıtput unit, and they are all 1.0 between Input Units 20–28 and the "above" unit. (A different sign should have been included for the coefficients for Input Units 1–14 and for Input Units 15–28; although the particular sign is arbitrary [it depends on how 0 and 1 were used to code the input], the above and below judgments should have opposite signs.[4]) Such information about the training patterns indicates that the networks were trained to treat Input Units 1–9 as encoding locations above those encoded by Input Units 20–28.

It is important to note that in Cook et al.'s (1995) scrambled input vectors, Input Units 7–10 are switched with Input Units 25–28, but the correlations between these units and "above" responses remain the same as before the units were switched (Cook et al., p. 412). Hence, after scrambling, these correlations still reveal that the networks were trained to treat Input Units 7–10 as encoding locations that are functionally above those encoded by Input Units 25–28. That is why Cook et al. found the same pattern of results for networks trained with our original vectors and for those trained with the scrambled vectors—nothing was functionally changed by scrambling the vectors (cf. Scalettar & Zee, 1988). Scrambling the original input array did not change the functionally defined locations of input units because the input–output mappings were not changed as well. From an outside observer's point of view the input was scrambled, but from the network's point of view a spatial representation was still present. Altering the input (e.g., scrambling) need not affect the functional relations if the mechanism that accesses the input compensates appropriately.

Thus, scrambling the input array does not guarantee that the network will respect the organization the modeler intends; the modeler must specify what is intended by the way the mapping between input and output is defined. Networks of the sort we used are not instructed in advance exactly how to map the input onto the output; they develop an appropriate mapping in the course of training. If it is advantageous to interpret the input spatially to produce the requisite output, the network may well do so.

Our claim that scrambling the input did not affect the spatial properties of the input representation may seem incompatible with the fact that many of the early cortical areas that process visual input are topographically organized (for a review, see Kosslyn, 1994). However, the topographically organized regions of the brain are preconfigured to perform the mappings a specific way, and thus the effects of scrambling would have to override a preexisting organization; this is not analogous to training a scrambled net from scratch. Cook et al.'s (1995) manipulation may at first glance seem compelling because one assumes that the input array has an inherent spatial organization, which is disrupted by scrambling—but the array does not have such an inherent spatial organization; the spatial organization only develops as the input–output mapping is established. If networks were trained with one mapping and then scrambled, they would probably not perform as well as unscrambled networks.[5]

thus make correct judgments that have no relation to human performance" (p. 412). It is an error to pit spatial information against specific combinations of on–off units. Putnam (1973) eloquently pointed out the distinction between explanations and "parents of explanations." If one wants to know why a square peg will not fit in a round hole, one does not ask about properties of molecules—one asks about shape and rigidity. There is a clear relationship between the two levels of analysis: Properties of molecules obviously contribute to properties of objects. But just as properties of objects cannot be replaced by properties of molecules, or properties of architectural styles cannot be replaced by properties of building materials, properties of representations and computations cannot be completely replaced by properties of individual units and connections. In networks, spatial information (or any kind of information) is specified by using combinations of on–off units, but it also can be characterized at a higher level of analysis. Depending on what question is being asked, different levels of analysis are appropriate for the answer; given that we are asking about spatial relations encoding, the level of representations and computations clearly is appropriate.

[4] Note also that correlations between the input units and only one of the two relevant output units are reported in each column of Table 3 in Cook et al. (1995). For example, the 10th column in Table 3 illustrates the product–moment correlation coefficients between input units and the "above" categorical output unit across the 80 training vectors used to replicate Study 1, Part 1 of Kosslyn et al. (1992). Because the coefficients associated with Input Units 15–28 are negative, these coefficients reflect correlations with the "above" output unit; activation of input units in the lower visual field tend to be negatively correlated with an above response. However, in Table 3 Cook et al. failed to provide the correlations with the "below" output unit. These correlations are critical because definitive information must be registered by high correlations with the opposite signs in the responses of the two output units; high correlations with the responses of both output units would not indicate definitive information any more than would low correlations with both output units, given that correct responses were always coded through two output units with opposite values (e.g., above unit on, below unit off). The correlations with the below output unit also indicate why the minus signs that are missing in the coefficients of Cook et al.'s Table 3 are important; without them, the network cannot recover the specific response.

[5] Moreover, even in the brain the topographically mapped areas do not preserve the image on the retina perfectly: The cortical representations typically expand the foveal regions, represent only a portion of the field, and are neither isotropic nor homogeneous. Thus, space is not veridically represented by the physical organization of these areas, and spatial properties are preserved only because of the nature of the connections (i.e., the mapping) to later areas. For spatial encoding, the mappings from these areas project to the posterior parietal lobe, which is not topographically organized in monkeys (Andersen, Asanuma, Essick, & Siegel, 1990). There is no qualitative difference between the kind of compensations that are necessary in these mappings and those that would be necessary if the lower level areas were not topographically orga-

However, even if there is a preexisting representation that is spatially organized, scrambling the input to it does not necessarily prevent the resulting representation from being spatial. Consider the results of two experiments that "scramble" the input of organisms. Sperry (1943) cut the optic nerve in newts and rotated the eye 180°. After the nerve regenerated, the newts behaved as if their visual world had been inverted and shifted left for right. Furthermore, no amount of practice could reverse the deficit. Different results were found in an experiment by Knudsen and Brainard (1991). These investigators raised barn owls with prismatic spectacles that exposed distinct portions of the retinae to relatively constant conditions of displaced vision, blocked vision, or normal vision. Where the lenses displaced visual input, large adaptive changes occurred in auditory spatial tuning in the tectal auditory space map; learning processes were able to compensate for the fact that the prisms gave the owls scrambled input. From the Sperry (1943) and Knudsen and Brainard (1991) experiments, can we conclude that newts, which could not learn to compensate for scrambled input, use spatial information, whereas owls, which can learn to compensate for scrambled input, do not use spatial information? On the basis of the logic they have applied in their critique of the Kosslyn et al. (1992) models, it would appear that Cook et al. (1995) would take the fact that owls can learn to compensate for scrambled input to imply that they do not use spatial information.

In short, then, whether a representation is spatial can only be determined within the context of a processing system; if a representation implicitly specifies geometric relations that can be used by other parts of a system, then that representation is spatial.

Receptive Fields

The fact that the networks spontaneously develop spatially localized receptive fields is evidence that the input had spatial properties. For a neuron, a receptive field is the set of locations in space where a stimulus will cause the neuron to respond; for network models, a receptive field for a hidden unit is the set of units in the input array that have high weights on the connections to that hidden unit (and hence input from those units has a relatively large influence on the activation state of the hidden unit). The Kosslyn et al. (1992) networks developed spatially localized receptive fields spontaneously, with high weights being defined over limited pockets of the input array for each hidden unit. In reference to Part 1 of Study 3 in Kosslyn et al., Cook et al. (1995) asserted "... this receptive field measure was in fact nothing more than another way of looking at the differences in net performance that were due to the *number* of input units with definitive information" (p. 421). But this is not so: The high weights were not distributed randomly over the input array; it was not the sheer number of such weights that was critical, but rather how they were organized. The receptive fields were organized spatially: Connections from nearby locations (with location being characterized within the context of the system itself) had high weights, and weights decreased with distance (defined by number of units) from the center of a region.[6] Such local organization is an inherent aspect of a spatial structure and would make no sense if the input array were treated in the model as an unstructured montage.[7]

According to Cook et al. (1995), "the main error in the original work was to conceptualize the tasks given to simple back-propagation networks as 'categorical' and 'coordinate spatial' without checking to see how the nets actually performed" (p. 410).[8] In fact, Kosslyn et al. (1992) and Jacobs and Kosslyn (1994) not only reported detailed analyses of receptive fields in order to discover the nature of the input–output mappings but also experimentally manipulated key features of these properties to confirm the original post hoc analyses and inferences. The major inference was that categorical spatial relations are consistently encoded more effectively when input units have relatively small, nonoverlapping receptive fields, whereas coordinate (metric) spatial relations are consistently encoded more effectively when

nized. The physically topographic organization appears to foster local inhibitory interactions, which are useful for detecting edges and similar properties. But such organization apparently is not critical for encoding spatial relations per se; if it were, the various distortions in the maps would affect the encoding of spatial relations, but they do not.

[6] This characterization of receptive fields in networks also lies at the heart of the models reported by O'Reilly, Kosslyn, Marsolek, and Chabris (1990). They provided networks with two-dimensional input arrays that contained a single point, and required the networks to indicate the X and Y coordinates of the point. These networks were hardwired to have large receptive fields, and they performed best when these receptive fields were distributed across the input array. This result is analogous to what was found in the Kosslyn et al. (1992) simulations of metric spatial relations encoding (Part 2 of Study 3 and Study 4). In both models, it was properties of the spatial structure of the input that dictated how well the models could encode location, and in both cases location was functionally defined.

[7] Although it is logically possible to register a categorical spatial relation with just a single input unit if the stimulus appears within a fixed region of the input field, we did not find such cases; in fact, the average receptive field size for categorical encoding was 4.8 units (Kosslyn et al., 1992, p. 570). It is also worth noting that although the simulations of metric encoding had fewer input units with definitive information, those units were distributed over a larger region than the input units in the simulations of categorical encoding.

[8] The appropriate terminology is *categorical spatial relations* and *coordinate spatial relations*—both types of representations are spatial. Cook et al. (1995) also apparently misread our earlier papers in other ways; for example, they claim that the Kosslyn (1987) snowball hypothesis of the development of hemispheric specialization "contrasts with the possibility that hemispheric *interactions* may also contribute significantly to functional differences. . . . It could be that their . . . mutual inhibitory influences . . . amplify small intrinsic differences" (Cook et al., p. 416). In fact, Kosslyn, Sokolov, and Chen (1989), who implemented a version of the Kosslyn (1987) theory, explicitly simulated the effect of mutual inhibition between the hemispheres on specialization.

input units have relatively large, overlapping receptive fields. For coordinate judgments, the networks relied on *coarse codes,* whereas for categorical judgments they did not. A coarse code corresponds to the pattern of output from a set of broadly tuned units (as occurs when units have large, overlapping receptive fields); the profile of relative contributions of each unit comprises the representation, in just the way that the relative outputs of the three types of cones on the retina specify hue. The importance of this aspect of network design is a computational finding that provides evidence for the conceptual distinction between the two types of encoding.

In summary, we find no grounds for accepting Cook et al.'s (1995) claim that the Kosslyn et al. (1992) networks did not use spatial information. Not only are spatial properties of input defined by the way the input is processed but also the existence of well-defined receptive fields is good evidence that the input in our models did function spatially.

Definitive Information Versus Spatial Relations

The second major issue raised by Cook et al. (1995) is independent of the first, and, in our view, poses a more serious challenge to the claim that Kosslyn et al.'s (1992) models provided an independent source of convergent evidence for the distinction between two types of spatial relations representation. Cook et al. argued that the purported differences between types of spatial relations representations were actually due to imbalances in the correlations between input and output unit activity, which rendered the results an artifact of differences in difficulty. Cook et al. correctly noticed that "... when any of the first 9 input units were on, the categorical judgment was *always* above (and when any of the last 9 units were on, *always* below)" (p. 417). Cook et al. argued that such definitive information allowed the networks to compute above and below without needing to compute spatial relations. We were interested in understanding spatial relations, that is, the assessment of one object's location relative to that of another.

Cook et al. (1995) have discovered a confound in the original Kosslyn et al. (1992) simulations, which leads us to view those results with caution. However, their observation indicates that another account is possible, not that the original one is necessarily incorrect.[9] Indeed, there is good reason to believe that Cook et al.'s possible account has limited generality and does not significantly undermine the Kosslyn et al. simulation results as evidence for two types of spatial relations computations. Accordingly, we next consider split versus unsplit networks, the distinction between definitive versus diagnostic information, and the relative importance of variations in receptive field size and definitive information.

Split Versus Unsplit Networks

Cook et al. (1995) focused on the first models reported by Kosslyn et al. (1992), in which two types of networks were compared. In some, a single network encoded both categorical and coordinate spatial relations; in others, the hidden layer of the network was segregated, so that different output streams were created. This technique was originally used by Rueckl, Cave, and Kosslyn (1989) to show that "what" and "where" are encoded best by a system that segregates the two types of information, as in fact occurs in the brain (e.g., Ungerleider & Mishkin, 1982). Cook et al. correctly noted that the mere fact that split networks can encode categorical and coordinate information more effectively than do unified networks is not strong evidence for the psychological validity of the distinction. They pointed out that this result merely demonstrates that two different types of mappings are performed when these two judgments are made by the models. However, Cook et al. went a step beyond this: They argued that the two judgments are not based on encoding categorical versus coordinate spatial representations but rather are based on distinguishing between large-amount-of-definitive-information versus small-amount-of-definitive-information. According to Cook et al., split networks perform better than unsplit networks because the two tasks differ in this regard, not because of anything having to do with encoding categorical and coordinate representations per se.[10]

We agree that one should not rely solely on the split network technique when using computational models to produce evidence that two processes are computationally distinct. However, if two processes do in fact rely on distinct computations, then their input–output mappings will interfere within a single network—and hence we expect a unified network to perform worse than a split network (provided that the hidden units are allocated appropriately; see Kosslyn et al., 1992, p. 566; Rueckl et al., 1989). The fact that another account for these particular simulations is possible does not completely undercut their utility: If the results had come out differently, with either no difference or with the unified network's performing better (as Kosslyn et

[9] In fact, the theory that was evaluated with the simulations also led us to predict that the hemispheres typically differ in the size of the regions of space that are monitored, and Kosslyn, Anderson, et al. (1994) later confirmed this prediction (but also showed that such differences can be modified by attentional factors). In addition, the claim that categorical spatial relations are processed more efficiently with smaller receptive fields predicts that blurring the stimuli should impair performance of a categorical spatial relations task more than a coordinate spatial relations task; Cowin and Hellige (1994) have demonstrated exactly this effect. Moreover, the fact that the models demonstrated effects of discriminability, which were based on the ease of defining the bins of space that are encoded, allowed Kosslyn et al. (1992) to explain effects of discriminability that were found in people (see Kosslyn, Koenig, et al., 1989).

[10] We have sometimes emphasized the distinction between processes that encode categorical versus coordinate information and sometimes emphasized the distinction between categorical versus coordinate spatial relations representations. We assume that distinct processes are used to encode the qualitatively distinct types of representations. Indeed, in networks the pattern of weights distributed on connections (the representation) actually delineates the course of processing.

al. showed that it does when two coordinate spatial relations must be encoded at the same time), this would have been evidence against the theory.

Cook et al.'s (1995) observations underline a critical aspect of our approach: The use of converging evidence. Just as we do not rely on a single type of evidence within the research program as a whole (relying on results from divided-visual-field studies, brain imaging, brain damage, and computational models), we do not rely on a single type of model to produce computational evidence. Indeed, Kosslyn et al. emphasized not the split–unsplit models, which Cook et al. have focused on, but rather the models in which receptive field sizes were manipulated or measured. These models make the closest contact with the novel aspects of the theory that have been subjected to successful empirical test (e.g., Cowin & Hellige, 1994; Kosslyn, Anderson, Hillger, & Hamilton, 1994).

Definitive Versus Diagnostic Information

Kosslyn et al. (1992) and Jacobs & Kosslyn (1994) have argued that when humans encode categorical spatial relations such as above–below and left–right, they define spatial "bins" that are used to perform the task. In fact, as Kosslyn et al. pointed out (p. 569), when it is difficult to delineate such bins, the left hemisphere should not have an advantage in this type of encoding; this observation allowed Kosslyn et al. to explain some of the otherwise paradoxical results of Sergent (1991). It is useful to distinguish between *definitive* information, which refers to the use of absolute locations instead of spatial relations, and *diagnostic* information, which merely signals a specific spatial relation. (By the term *absolute,* we mean a fixed location within a reference field, such as an array in a computer or on a screen in front of a subject.) Kosslyn et al. (1992) stated the following:

> Sets of these [small] receptive fields would define particular areas, which could be used to specify regions that are above or below a reference point . . . if the receptive fields did not overlap at all and the categories corresponded to discrete regions of space, such mappings would be 'linearly separable'—so straightforward that they could be accomplished by direct connections from the input units to the output units. . . . (p. 569)

In other words, in some cases the presence of stimuli at two locations can directly signal a specific categorical spatial relation. Note that this is a true spatial relation: What is critical is the relative positions, not the absolute position.

The difference between definitive and diagnostic information can be illustrated by contrasting two versions of the "analogous psychological experiment" offered by Cook et al. (1995). They described a situation in which subjects must determine whether degraded images of faces are male or female and are young or old. In their analogy, 24 of the 40 male faces have a blob on the left and 24 of the 40 female faces have a blob on the right; also, 6 of the male and 6 of the female faces have a blob on the top, which always indicates that the face is young. Their point is that the blobs alone would provide definitive cues that allow subjects to make the judgments without actually processing the faces. In this case, the blobs are added to the stimuli, they are not intrinsic to them. Now consider an illustration of diagnostic information: The size of the eyebrow ridge is large for 24 of the 40 male faces and short for 24 of the 40 female faces, and wrinkles appear on 6 of the older male faces and 6 of the older female faces. In these cases, the cues are an inherent part of the stimuli, not something extraneous that is added later.

Our claim is not that categorical spatial relations are encoded by seeking the presence of a cue at a specific place in the visual field; rather, by encoding the presence of stimuli at two specific locations, one can quickly encode some categorical spatial relations. Such cues are not always effective, however; Ullman (1985) argued that in some cases on–off or inside–outside judgments can be made only after a boundary is traced (for evidence, see Jolicoeur, Ullman, & Mackay, 1986). Even in such cases, however, it would still be useful to monitor outputs from relatively small receptive fields, and the categorical spatial relations encoding subsystem should play a role in adjusting the attentional scale and comparing relative locations.

Receptive Field Size Versus Definitive Information

As Cook et al. (1995) pointed out (see p. 422), by hardwiring receptive fields of different sizes, we greatly reduced the correlations between individual input and output units.[11] Nevertheless, we still observed clear and principled differences between categorical and coordinate encoding (e.g., Kosslyn et al., 1992, Study 3, Part 2). The results from the hardwired simulations lined up well with those found when networks spontaneously developed similar receptive fields. The convergent findings when the imbalances are small and when they are large show that such imbalances alone cannot explain all of the processing differences between the two tasks.

In addition, there are analytical reasons for concluding that units with small, nonoverlapping receptive fields are best for encoding categorical visual representations,

[11] It is difficult to know what to make of Table 4 in Cook et al. (1995). They claim to be discussing the receptive fields added by Kosslyn et al. (1992) in their Study 3, Part 2 and Study 4 (see Cook et al., p. 421). Thus, they provide point-biserial correlation coefficients between input units with "large" receptive fields and the output units, presumably computed across all 80 training patterns. However, whereas Cook et al. gave values for 28 input units, Kosslyn et al. used just 14 receptive fields in the crucial experiment (Study 3, Part 2) that produced the expected categorical–coordinate dissociation. When we analyzed the receptive field-filtered input arrays actually used by Kosslyn et al. (pp. 570–572), we found mean correlations of .19 and .38 for coordinate and categorical tasks, respectively, with large receptive fields (σ = 1.42), much lower than the .26 and .57 means reported by Cook et al. (Furthermore, although Cook et al. also showed correlations for large receptive fields in the dual-coordinate task used in Study 1, Part 2 of Kosslyn et al., this task was not actually used in receptive field simulations by Kosslyn et al.)

whereas units with large, overlapping receptive fields are best for encoding coordinate representations. Hinton (1981) provided a quantitative understanding of the relationship between receptive field size and resolution in the case of a set of binary units, each of which becomes active when a stimulus falls within its receptive field. Let D denote the d.ameter of a unit's receptive field, k denote the dimensionality of the space to be represented, and N denote the desired number of just-noticeable differences in each dimension (i.e., the desired resolution). The number of units required to achieve this resolution is N^k/D^{k-1}. That is, for a fixed number of units, a high-resolution coarse code (i.e., a code with a large N) requires units with large receptive fields (fields with a large diameter D), whereas a low-resolution code can be achieved with units with small receptive fields. (Note that this equation applies only when the input array has two or more dimensions; Jacobs & Kosslyn, 1994, essentially replicated the Kosslyn et al., 1992, results when two-dimensional stimuli were used.)

If we make the reasonable assumption that coordinate visual tasks require relatively high spatial resolution, then this analytic result implies that filtering input through units with large receptive fields is especially good for performing these judgments. In contrast, filtering input through small, nonoverlapping receptive fields should facilitate encoding categorical spatial relations for reasons noted above. For present purposes, it is of interest that the relationships between task, resolution, and receptive field size are independent of the presence or absence of correlations between input and output units; neither the equation governing the relationship between resolution and receptive field size nor the assumptions regarding the relationships between task and receptive field size make use of such correlations. Thus, our conclusions are the same whether or not definitive information is available in a particular categorical or coordinate task.

Finally, Jacobs and Kosslyn (1994) reported results that document the independence of the relationship between effect of receptive field sizes and the presence or absence of correlations among input and output units. In their network models, Jacobs and Kosslyn placed two stimuli at different locations on a two-dimensional input array; one stimulus was a particular shape and the other was a horizontal bar. The networks encoded not only spatial relations but also whether the stimulus was a specific object or a member of a shape category. Fourteen Gaussian units with specific receptive field sizes registered input in the array. That is, each pattern of activation over the input units was transformed to a pattern of activation over the Gaussian units; this transformed pattern may be thought of as a representation that corresponds to a point in a 14-dimensional space, which we call *Gaussian unit space*. Note that Gaussian units with different receptive field sizes represent a given stimulus on the input array as different points in this space.

Jacobs and Kosslyn (1994) defined a measure of the "goodness" of such representations as the minimum Euclidean distance between points in Gaussian unit space for different judgments (e.g., above versus below). If this distance is small, then the nearest points from different judgments are difficult to distinguish. In contrast, if this distance is large, then the nearest points from different judgments are easily distinguished. The results showed that Gaussian units with small, nonoverlapping receptive fields provided a better representation for encoding categorical spatial relations and shape categories, whereas Gaussian units with large, overlapping receptive fields provided a better representation for encoding coordinate spatial relations and specific shape exemplars. This result cannot depend on the presence or absence of correlations between input units and output units: These correlations were the same for representations derived from input units with different receptive field sizes.

In summary, the presence of definitive information in the input should make us cautious in interpreting the results from the Kosslyn et al. (1992) split–unsplit network simulations; the available results do not allow us to discriminate between the two interpretations. However, such definitive information does not undercut the fact that different-sized receptive fields were more effective for encoding the different judgments, which stands as computational evidence for the proposed distinction between two types of spatial relations encoding.

Conclusion

Computational modeling in cognitive neuroscience is a tool that can help us understand how the human brain accomplishes certain functions. As such it is an integral part of a larger research program that frames the issues to be modeled and uses the results of modeling as impetus for additional experimentation. Our analysis and modeling suggested to us that the visual system encodes categorical spatial relations, such as above–below, in part by carving space into discrete bins, which is accomplished effectively by using outputs from units with relatively small receptive fields—once different portions of a stimulus are localized in specific bins, the judgment can be made easily. In contrast, we found evidence that metric spatial relations are encoded via coarse coding, which depends on the comparisons of outputs from units with large, overlapping receptive fields.

Cook et al. (1995) offered two major criticisms of our computational models. The first, the claim that the models did not speak to how spatial relations are encoded, is off the mark. What counts as "spatial" can only be assessed from within the context of a processing system; if a representation implicitly specifies geometrical relations (from the point of view of processes that use that representation), it is spatial. The second criticism has more force. Cook et al. observed that the presence of a stimulus in some regions of the input array could signal the correct response, and thus the networks need not have computed categorical spatial relations. This criticism leads to a viable (but not necessarily correct) alternative interpretation of the results from split and unsplit networks, but does not apply to the finding that different-sized receptive fields facilitate the two types of spatial relations encoding; this form of computational evidence is critical because it is tied closely to recent (sometimes counterintuitive) empirical results with humans (e.g., Cowin & Hellige, 1994; Kosslyn, Anderson, et al., 1994).

We agree that it is important to explore further the role of such distinctive information in these kinds of models. One way to do so would be to use large-scale, realistic training sets in which there are many objects, each of which projects many different patterns of "light" on the input array because of translation, rotation, changes in illumination and so on; in such networks definitive information would be not available in large portions of the input array.

In closing, we wish to emphasize that computational models of human performance can not only provide support for a conceptual distinction but also lead one to conduct new, fruitful empirical research with humans. For a model to be useful, one must not only understand how it operates but also must derive specific predictions from its properties. Cook et al. (1995) summarized their objections to Kosslyn et al. (1992) by stating "Their recent work . . . has not been successful. . . . The effects do not correspond to human performance on similar tasks. . . ." (p. 422). We feel that this is an overstatement, given that the simulations were designed to perform one of the tasks Hellige and Michimata (1989) and Kosslyn, Koenig, et al. (1989) administered to humans, and that the models did indeed capture many features of human performance on those tasks (such as the effects of discriminability on performance; see Kosslyn et al., Study 2). The models were used in conjunction with research on how human subjects encode spatial relations; the primary measures of their adequacy are how well they fit the prior data and the extent to which their predictions are later confirmed empirically.

In our view, network models of the sort we implemented are but one form of converging evidence. Such evidence is useful in combination with results from behavioral experiments with normal subjects, tests of patients with brain damage, and findings from brain-scanning techniques. Computational models also can help one to account for previously reported data and to derive predictions for new studies with human subjects. Our models have played these roles well.

References

Andersen, R. A., Asanuma, C., Essick, G. K., & Siegel, R. M. (1990). Cortico-cortical connections of anatomically and physiologically defined subdivisions within the inferior parietal lobule. *Journal of Comparative Neurology, 296,* 65–113.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain.* Cambridge, MA: MIT Press.

Cook, N. D., Früh, H., & Landis, T. (1995). The cerebral hemispheres and neural network simulations: Design considerations. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 410–423.

Cowin, E. L., & Hellige, J. B. (1994). Categorical versus coordinate spatial processing: Effects of blurring and hemispheric asymmetry. *Journal of Cognitive Neuroscience, 6,* 156–164.

Hellige, J. B., Bloch, M. I., Cowin, E. L., Eng, T. L., Eviatar, Z., & Sergent, V. (1994). Individual variation in hemispheric asymmetry: Multitask study of effects related to handedness and sex. *Journal of Experimental Psychology: General, 123,* 235–256.

Hellige, J. B., & Michimata, C. (1989). Categorization versus distance: Hemispheric differences for processing spatial information. *Memory & Cognition, 17,* 770–776.

Hesse, M. B. (1963). *Models and analogies in science.* London: Sheed and Ward.

Hinton, G. E. (1981). Shape recognition in parallel systems. In A. Drina (Ed.), *Proceedings of the Seventh International Joint Conference on Artificial Intelligence.*

Jacobs, R., & Kosslyn, S. M. (1994). Encoding shape and spatial relations: The role of receptive field size in coordinating complementary representations. *Cognitive Science, 18,* 361–386.

Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition, 14,* 129–140.

Knudsen, E. I., & Brainard, M. S. (1991). Visual instruction of the neural map of auditory space in the developing optic tectum. *Science, 253,* 85–87.

Kosslyn, S. M. (1980). *Image and mind.* Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1987). Seeing and imagining in the cerebral hemispheres: A computational approach. *Psychological Review, 94,* 148–175.

Kosslyn, S. M. (1994). *Image and brain.* Cambridge, MA: MIT Press.

Kosslyn, S. M., Anderson, A. K., Hillger, L. A., & Hamilton, S. (1994). Hemispheric differences in sizes of receptive fields or attentional biases? *Neuropsychology, 8,* 139–147.

Kosslyn, S. M., Chabris, C. F., Marsolek, C. J., & Koenig, O. (1992). Categorical versus coordinate spatial relations: Computational analyses and computer simulations. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 562–577.

Kosslyn, S. M., Gitelman, D., Thompson, W. L., Rauch, S. L., & Alpert, N. M. (1994). Encoding categorical and coordinate spatial relations representations: A PET study. Manuscript in preparation.

Kosslyn, S. M., & Koenig, O. (1992). *Wet mind: The new cognitive neuroscience.* New York: The Free Press.

Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., & Gabrieli, J. D. E. (1989). Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 723–735.

Kosslyn, S. M., Sokolov, M. A., & Chen, J. C. (1989). The lateralization of BRIAN: A computational theory and model of visual hemispheric specialization. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert Simon* (pp. 3–29). Hillsdale, NJ: Erlbaum.

Laeng, B. (1994). Lateralization of categorical and coordinate spatial functions: A study of unilateral stroke patients. *Journal of Cognitive Neuroscience, 6,* 189–203.

Laeng, B., & Peters, M. (in press). Cerebral lateralization for the processing of spatial coordinates and categories in left- and right-handers. *Neuropsychologia.*

Marshall, J. C. (1981). Hemispheric specialization: What, how, and why. *Behavioral and Brain Sciences, 4,* 72–73.

Nebes, R. D. (1978). Direct examination of cognitive function in the right and left hemispheres. In M. Kinsbourne (Ed.), *Asymmetrical function of the brain.* Cambridge: Cambridge University Press.

O'Reilly, R. C., Kosslyn, S. M., Marsolek, C. J., & Chabris, C. F. (1990). Receptive field characteristics that allow parietal lobe neurons to encode spatial properties of visual input: A computational analysis. *Journal of Cognitive Neuroscience, 2,* 141–155.

Putnam, H. (1973). Reductionism and the nature of psychology. *Cognition, 2,* 131–146.

Rueckl, J. G., Cave, K. R., & Kosslyn, S. M. (1989). Why are "what" and "where" processed by separate cortical visual systems? A computational investigation. *Journal of Cognitive Neuroscience, 1,* 171–186.

Scalettar, R., & Zee, A. (1988). Perception of left and right by a feed forward net. *Biological Cybernetics, 58,* 193–201.

Sergent, J. (1991). Judgments of relative position and distance on representations of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance, 91,* 762–780.

Sperry, R. W. (1943). Effect of 180 degree rotation of the retinal field on visuomotor coordination. *Journal of Experimental Zoology, 92,* 263–279.

Ullman, S. (1985). Visual routines. In S. Pinker (Ed.), *Visual cognition* (pp. 97–159). Cambridge, MA: MIT Press.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *The analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Received May 6, 1994
Revision received November 8, 1994
Accepted November 30, 1994 ■