

# **Things and Places: How the Mind Connects With the World**

Zenon W. Pylyshyn  
Rutgers Center for Cognitive Science

*Forthcoming, 2007, MIT Press (Jean Nicod Lecture Series)*

## Table of contents

### Preface & Acknowledgements

#### Chapter 1. Introduction to the Problem: Connecting Perception and the World

- 1.1 Background
- 1.2 What's the problem of connecting the mind with the world?  
Doesn't every computational theory of vision do that ?
- 1.3 The need for a direct way of referring to certain individual tokens in a scene
  - 1.3.1 *Incremental construction of representations (and a brief sketch of FINSTs)*
  - 1.3.2 *Using descriptions to pick out individuals*
  - 1.3.3 *The need for demonstrative reference in perception*
- 1.4 Some empirical phenomena illustrating the role of Indexes
  - 1.4.1 *Tagging/marking individual objects for attentional priority*
  - 1.4.2 *Argument binding*
  - 1.4.3 *Subitizing*
  - 1.4.4 *Subset selection*
- 1.5 What are we to make of such empirical demonstrations?

#### Chapter 2. Indexing and tracking individuals

- 2.1 Individuating and tracking
- 2.2 Indexes and primitive tracking
- 2.3 What goes on in MOT?
  - 2.3.1 *FINSTs and Object Files*
  - 2.3.2 *The explanation of tracking*
- 2.4 Other empirical and theoretical issues surrounding MOT
  - 2.4.1 *Do we track by keeping a record of locations?*
  - 2.4.2 *Can we select objects voluntarily?*
  - 2.4.3 *Tracking without keeping track of labels*
  - 2.4.4 *Nonconceptual individuation without reference?*
- 2.5 Infants capacity for individuating and tracking objects
- 2.6 Summary and Implications for the Foundations of Cognitive Science
  - 2.6.1 *Review: Nonconceptual functions and Natural Constraints*
  - 2.6.2 *Summary: Why are FINSTs needed?*

#### Chapter 3. Selection: The key to linking representations and things

- 3.1 Selection: The role of focal attention
  - 3.1.1 *Allocating and shifting attention: The role of objects and places*
  - 3.1.2 *Attention selects and adheres to objects*
- 3.2 Selection and demonstrative reference: The role of FINSTs
  - 3.2.1 *Causes and codes*
  - 3.2.2 *Conceptual and nonconceptual representations and quasi-representations*
- 3.3 Problems with selection-by-location
  - 3.3.1 *A note on the role of location in selection and tracking*
  - 3.3.2 *Selection and the binding problem*
  - 3.3.3 *Selection and the causal link*
  - 3.3.4 *Selection as nonconceptual access*
- 3.4 Feature placing and sentience
- 3.5 What do FINSTs select? Some consequences of the present view

## **Chapter 4. Conscious contents and nonconceptual representation**

- 4.1 Nonconceptual representation and perceptual beliefs
- 4.2 The role of conscious experience in the study of perception and cognition
  - 4.2.1 *The Contents question*
  - 4.2.2 *Conscious experience and public report*
- 4.3 What subjective experience reveals about psychological processes
  - 4.3.1 *The illusion of conscious will*
  - 4.3.2 *Conscious Experience, Interpretation and Confabulation*
  - 4.3.3 *Failures of Conscious Access: Split Brains and Split Visual Systems*
- 4.4 The phenomenal experience of seeing
  - 4.4.1 *Nonconceptual representation in visual perception*
- 4.5 The phenomenal experience of mental imagery
  - 4.5.1 *Phenomenal experience and explanation: The role of tacit knowledge*
  - 4.5.2 *Does the architecture of visual cortex matter to explanations of imagery?*
  - 4.5.3 *Problems in accounting for phenomenal space by appealing to brain-space*
- 4.6 Does phenomenal appearance correspond to a level of representation?

## **Chapter 5. How we represent space: Internal vs External Constraints**

- 5.1 What does it mean to represent space?
- 5.2 Internalizing general spatial constraints
  - 5.2.1 *Marr's principle of Natural Constraints*
  - 5.2.2 *Shepard's psychophysical complementarity*
- 5.3 Internalizing spatial properties by mapping them onto an inner space
  - 5.3.1 *Brain Space*
  - 5.3.2 *Functional space and principled constraints*
  - 5.3.3 *Internalizing by incorporating visuomotor experience: Poincaré's insights*
- 5.4 What is special about representing space?
  - 5.4.1 *Conditions on a theory of Active Spatial Representation (ASPAR)*
  - 5.4.2 *Maps and map-like navigation behavior*
- 5.5 Externalizing spatial properties: The projection hypothesis
  - 5.5.1 *Where is our spatial representation?*
  - 5.5.2 *FINST Indexes can anchor spatial representations to real space*
- 5.6 Projection in non-visual modalities
  - 5.6.1 *The unitary frame of reference assumption*
  - 5.6.2 *The role of coordinate transformations in externalizing space*
  - 5.6.3 *Proprioceptive FINSTs? The projection hypothesis in nonvisual modalities*
  - 5.6.4 *Motor control, proprioception and intention define places in allocentric space*
  - 5.6.5 *Summary of evidence in support of the projection hypothesis*

Conclusions

References

## Preface & Acknowledgements

This book is based in part on the Jean Nicod Lectures that I delivered in Paris in May-June of 2004. The temporal gap between the lectures and the publication is not due entirely to my slow typing, but arose from my need to assimilate the rather wide range of publications that are relevant to the thesis I am presenting. The thesis, it turns out, is one that I have been gestating over many years and hints of it occur in fragmentary form in a number of my publications. Many of these are reports of experimental work carried out with graduate students over the years and whose contribution is much appreciated.

The thesis rests on a growing appreciation of an idea I first learned from David Marr who refers to it as the principle of Natural Constraints. The mind has been tuned over its evolutionary history so that it carries out certain functions in a modular fashion, without regard for what an organism knows or believes or desires, but because it is in its nature, or as I more often put it, because of its architecture. So far this is an innocent enough idea that fits many different schools of psychology (and in fact is at the heart of J.J.Gibson's Direct Realism Theory). The particular constraint I am interested in here takes the form of a mechanism that allows the modular perceptual system to do things that many philosophers have said (correctly) can only be done by using the sophisticated machinery of concepts and the logical machinery of induction, deduction and what Charles Sander Peirce called abduction. The mechanism includes the capacity to select individual things in one's field of view, to re-identify each of them under certain conditions as the same individual thing that was seen before and to keep track of their enduring individuality despite radical changes in their properties. I claim that so long as we are in the kind of world we live in there are mechanisms that allow the visual system to do these things without using the heavy equipment of concepts, identity, and tenses (which really are needed in the general case). For example this is a world in which most surfaces that we can see are surfaces of physical objects, so that most of the texture elements we see move coherently as the object moves, almost all elements nearby on the proximal image are at the same distance from the viewer, and when objects disappear they frequently reappear nearby and often with a particular pattern of occlusion and disocclusion at the edges of the occluding opaque surfaces, and so on. Identifying things as ones we have seen before and keeping track of them as being the same individual things over time is at the heart of the research I have been doing and it has shown that we are very good at doing it in a way that does not use encoded properties (nor the conceptual category) of the things that are tracked and re-identified. This mechanism is important to us because if it were not for the existence of such nonconceptual processes, our concepts would not be grounded in experience so they would not have the meaning that they do. I have proposed that the capacity to individuate and track several independently-moving things is accomplished by a mechanism in the early vision module that I have called FINSTs (for historical reasons I called them "Fingers of INSTantiation" because they were initially viewed as a mechanism for instantiating or binding the arguments of visual predicates to objects in the world). This primitive nonconceptual mechanism functions to identify, reidentify and track distal objects. It is an ability that we exercise every waking minute, and has also been understood to be fundamental to the way we see and understand the world.

I came upon these ideas in quite a different context, initially one in which I (along with my colleague Edward Elcock) attempted to develop a computer system for reasoning about diagrams and later when I was carrying out experimental research on vision, visual attention and mental imagery. This may seem like a very circuitous route, but it has turned out that all these endeavors involved the same puzzles, which I later discovered were also the puzzles that preoccupied many philosophers: The puzzle of how concepts are grounded in experience, how

we manage to encode and represent properties of the world when there are so many of them, why we feel that we are conscious of seeing an enormous number of things but are unable to report most of them, while at the same time a great deal of information of which we are not conscious can affect our behavior. These puzzles appear in their most striking form in discussions of two related problems: What are properties of mental images that allows them to function in thought, and How do certain kinds of thoughts – thoughts about spatial layouts – manage to display properties very similar to those of perceived space.

In this book I examine a number of critical functions of early vision (the part of vision that is informationally encapsulated from the rest of the mind) in the light of the FINST mechanism. Chapter 1 looks at the nature of the problem that FINST are intended to solve, from our initial encounter with the problem of incrementally building a representation over time as various aspects are noticed, to the deep problem of grounding concepts in sensory information. This chapter also introduces FINST theory in terms of a number of experiments that illustrate their nature as pointers to things in the perceived world. It also offers the suggestion that FINSTs serve to provide what philosophers have called demonstrative reference or demonstrative identification. Chapter 2 focuses on a particular function that FINSTs serve – they allow nonconceptual *tracking* of individual things that move and change their properties. Since tracking is one of the critical aspects of our commerce with the world, these experiments serve as concrete examples of the role that FINSTs play in this process. It also provides a basis for a number of additional properties of this mind-world connection – it shows that things can be tracked as unidentified things with an enduring numerical identity (where by “unidentified” I mean they are not represented in terms of any conceptual category or in terms of distinctive properties). The notion of tracking also links this work to some findings in cognitive development where it has been shown that very young infants (under 6 months of age) are able to keep tracking of things that briefly disappear and are able to anticipate how many hidden things there are (as long as there are 3 or fewer).

Chapter 3 looks more closely at one of the functions that FINSTs perform – that of selecting things through something like attention (FINSTs are not exactly the same as focal attention and I discuss the differences). This chapter raises some of the controversial aspects of attentional selection. It argues that selecting is nonconceptual and does not depend on the prior encoding of any properties of the things selected – including their location. I argue that the reason this seems unintuitive is that we fail to distinguish between the various roles that the properties of things play in this process. Properties are involved in picking out things to which FINSTs are assigned, they are involved in determining whether things can be tracked, but they need not be encoded and used in the process of maintaining the identity of the things that are tracked. I spend time in Chapter 3 distinguishing between causes and codes and suggest that we should be conservative in describing certain mind-world connections are representations. This brings us to an important function that FINST selection plays, solving what has been called the *binding problem* – the problem of encoding certain sets of properties as being conjoined, as being properties of the same thing, as opposed to being properties that merely occur in the scene. While much of the psychological and philosophical literature sees the binding problem as being solved in terms of the co-location of properties, my proposal is that properties are considered conjoined if they are properties of the same FINSTed thing. This brings us to a point where we can say roughly what FINSTs attach to – what it is that I have been calling “things” (or even FINGs, to indicate that they are interdefined with FINSTs, as things that FINSTs select and refer to) – they attach to what in our sort of world typically turn out to be individual visible physical objects. I discuss the frequently cited notion of nonconceptual representation and suggest how this idea is closely related to the story I am telling about FINSTs. In Chapter 3 I propose that the only nonconceptual representational content we have is that secured by FINST indexes.

The view that only properties of FINSTed things get represented puts me in conflict with those who appeal to the richness of conscious experience in defending nonconceptual representation. For this reason I devote Chapter 4 to a discussion of the role of conscious contents in the process of connecting mind and world. What I end up claiming is that the contents of conscious experience are only one source of evidence for mental contents, and not even a very reliable one. I claim that there is no level of representation that corresponds specifically to the content of conscious experience and therefore that equating nonconceptual representation with the content of conscious experience is a mistake. The discussion of the content of conscious experience brings the topic around to the nature of the mental images that we experience when reasoning about concrete sensory appearances, about spatial layouts or when we use spatial mental model models in reasoning. In chapter 4 I focus primarily on what we can make of the contents of conscious experience and I use theories of mental imagery as the example of how we are misled when we view conscious content as a type of representation.

It is not until Chapter 5 that I focus directly on the problem of the representation of space in active working memory (as opposed to long-term memory). Here I review some of the proposals for how some mental representations manage to exhibit certain sorts of spatial properties. Most of these proposals hypothesize some internal constraints on representations of spatial layouts. The most widely accepted and intuitive proposal is that the spatial character of representations of space arise because the representations themselves are instantiated by spatial layouts in the brain – what I call neural layouts. I review this proposal as well as some related ones that attempt to retain the benefits of neural layouts without assuming a spatial arrangement in the brain. Of these, the one referred to as “functional space” fails to have any explanatory value and the other, which I call the spatial architecture proposal, reduces essentially to the literal space alternative. In this discussion I distinguish representations involved in long-term memory from representations I refer to as Active Spatial Representations (ASPAR), which are constructed both by vision and by reasoning that relies on imagined geometrical of spatial layouts. I then list what I take to be some of the conditions that an ASPAR must meet, which includes its capacity to represent magnitudes, its spatial configurational stability, its amodal nature, and its intimate connection with the motor system. The latter brings us to an overview of Poincaré’s proposal.

I finish Chapter 5 by presenting what might be seen as a fairly radical speculative proposal for an externalist theory of spatial representation in ASPARs (i.e., in spatial reasoning). The hypothesis, which arises from the ideas about FINSTs that I discuss in the first part of the book, is what I call the *Projection Hypothesis*. This proposal claims that in constructing a spatial representation from approximate, partial and qualitative information stored in long-term memory, what we do is pick out things in the concurrently perceived world using FINSTs and associate imagined objects with them (to think of the imagined objects as being located at the sensed objects). This allows us to use the perceptual system to draw inferences by pattern recognition rather than logical reasoning. I then generalize the projection hypothesis to nonvisual modalities such as proprioception, which requires that I deal with the multiple frame of reference problem (as opposed to a unitary allocentric frame of reference). In this task, coordinate transformation mechanisms, which are plentiful in the brain (especially in parietal cortex, as well as in superior colliculus and premotor cortical areas) play a central role.

Throughout this essay I try to draw morals for a number of philosophical issues such as whether there are nonconceptual representations, how concepts are grounded in perception, and how the mind deals with spatial properties. Clearly this palette is more than can be dealt with adequately in one short book. Yet there are some clear themes that run through these puzzles, especially when they are considered against the background of experimental findings in psychophysics, cognitive science, cognitive development, and neuroscience. So this is my

attempt to find a way through that forest by focusing on one or two sunny spots where I think progress has been made in the past two decades.

In this pursuit I must express my gratitude to the Institute Jean Nicod, who generously invited me to give these lectures, to the Institute Jean Nicod, the Centre Nationale de la Recherche Scientifique (CNRS), L'École des Hautes Études en Sciences Sociales (EHESS), and L'École Normale Supérieure (ENS) who funded the Jean Nicod Prize and provided space for me the following year as I worked on this book. In particular I wish to thank François Récanati, who chairs the Nicod Prize committee and looked after the logistics of my visit, Pierre Jacob, who directs the Institute Jean Nicod, and the many people from the Institute who made my stay memorable, especially Roberto Casati, Jérôme Dokic, Élisabeth Pacherie, Joelle Proust, and Dan Sperber. Michel Imbert and Sylvia Duchacek-Imbert were most welcoming and made our stay memorable.

I also acknowledge the help I received, intended or not, from my discussions with my friend Jerry Fodor and with colleagues Charles R. Gallistel (who helped to educate me out on the subject of navigation), Georges Rey (who carefully read an earlier version of the manuscript and tried his best to keep me from making some embarrassing philosophical slips and misusing some philosophical terms of art), to Alan Leslie and Lila Gleitman (with whom I have taught some of this material in joint courses), and to Susan Carey, Ned Block, Luca Bonatti. I am also grateful to the participants of the conference on Spatial Frames of Reference held in Paris in November 2005, including organizers Jérôme Dokic, Élisabeth Pacherie, and participants with whom I had a chance to try out the ideas in Chapter 5: Jean-René Duhamel, Yves Rossetti, Charles Spence, Barbara Tversky, Yann Coello, Paolo Bartolomeo and Sylvie Chokron.

## **Chapter 1. Introduction to the Problem: Connecting Perception and the World**

### **1.1 Background**

Just as Moliere's Monsieur Jourdain discovered that he had been speaking prose all his life without realizing it, so I discovered not too long ago that what I had been doing without realizing it occupies a position in the philosophical landscape. I discovered that coming from a very different perspective I had taken a position on a set of questions that philosophers had been worrying about for at least the past 50 years; questions about how concepts connect with the world, questions about whether there are nonconceptual representations and if so what they are like, as well as general questions concerning the grounding of mental states in causal connections with states of the world and, most recently, questions about how mental representations – such as those underlying mental imagery – attain their apparent spatial character that allows them to be used the way diagrams are used in thought. I propose, in this first chapter, to illustrate the questions that led me to work on these problems and then to describe, with the aid of some illustrative experiments, why there is a special problem of connecting representations with the world.

The central topic is the relation between the mind and the world. To a vision scientist this sounds like a strange topic. Isn't all of vision science about this? What's wrong with a story that begins with light falling on objects in the world and being reflected to the eye, where it is refracted and focused onto the retina, from which it is transformed into nerve impulses which encode various properties of the retinal stimulus, and transmit them to the visual cortex from where they are transformed once again, in ways that neuroscience is currently making good progress studying? Apart from a whole lot of missing details, it is of interest to ask what's missing from this general kind of story – isn't that what cognitive science and neuroscience are all about? Is there something missing in principle from this *kind* of story?

The answer I will offer is that there are important aspects of vision that such a story does not address. In this monograph I will attempt to describe some of what is missing and to illustrate the claims by describing relevant empirical research. The ideas come equally from philosophy, psychophysics and neuroscience.

### **1.2 What's the problem of connecting the mind with the world? Doesn't every computational theory of vision do that ?**

The basic problem is a familiar one in cognitive science: there are different levels of explanation, and different kinds of questions must be addressed in different vocabularies. The reason we need different vocabularies is that if the world is organized in certain ways there are different generalizations that can be captured in different vocabularies. Notwithstanding our belief in the unity of science, we do not address questions of economics or even of geology (which is concerned with such things as rivers and mountains) or meteorology (which is concerned with weather patterns) though theories of physics or chemistry, despite the fact that the tokens in all cases are physical. Let me illustrate the case for perception with a very simple example from the ancient history of vision research.

The goal of understanding what was regarded as humans' most noble sense has a long history, starting, as usual, with the ancient Greeks and taking a great leap forward in the late eighth century Arab world under al-Kindi, when the science of optics was brought into contact



with the study of visual perception. This path reached its peak with Johannes Kepler's brilliant solution, in the early 17<sup>th</sup> century, of the problem of the optics of the eye and his seminal recognition of the critical role that the retinal image plays in vision. But in the century that followed, this sudden spurt of progress seems to have gone into a hiatus. Kepler himself recognized that he had gone as far as he could with the set of concepts available to him. He wrote (quoted in Lindberg, 1976, p 202);

“I say that vision occurs when the image of the whole hemisphere of the world that is before the eye ... is fixed in the reddish white concave surface of the retina. How the image or picture is composed by the visual spirits that reside in the retina and the [optic] nerve, and whether it is made to appear before the soul or the tribunal of the visual faculty by a spirit within the hollows of the brain, or whether the visual faculty, like a magistrate sent by the soul, goes forth from the administrative chamber of the brain into the optic nerve and the retina to meet this image, as though descending to a lower court – I leave to be disputed by [others]. For the armament of the opticians does not take them beyond this first opaque wall encountered within the eye.”

In this quotation, Kepler touches upon a number of problems that are still active concerns, particularly the balance between top-down and bottom-up analysis of visual information (which he describes in terms of an administrative metaphor). But the quotation also provides a glimpse of Kepler's insightful acknowledgement that there remained serious problems that could not be addressed given the concepts of the day (“the armament of the optician”). What made Kepler particularly pessimistic is that, despite years of trying, he could find no way, within geometrical optics, to deal with the problem of the inverted and mirror-reversed image on the retina. This puzzle left a generation of brilliant mathematicians and thinkers completely stymied. Why? What did they lack? It is arguable that they lacked the abstract concept of *information* which did not come along fully until the 20th century. The concept of information made it natural to see *right-side up* and *upside down* as mere conventions, and allowed a certain barrier to be scaled because information only requires a consistent mapping and not the preservation of appearance. As (Dretske, 1981) points out, so long as the visual pattern is (nonaccidentally) correlated<sup>1</sup> with, and thereby carries information about some state of affairs, the information is then available to the right sort of processor which can, in principle, interpret it appropriately, taking into account how the information relates to subsequent uses to which it is put (e.g., object recognition and motor action). But even after we see that the information carried is the same in the right-side up as in the upside down image, there is still an obstacle at least as inscrutable as the one that held back Kepler; it is the gap between the incoming causally-linked information and *representational content*. If similarity of appearance is eliminated as a criterion, then what makes something a representation of a particular scene rather than of some other scene from which it could equally be mapped in a consistent (information-preserving) manner, and indeed, why are some states representations at all? This puzzle will occupy us throughout this book because its resolution (or

---

<sup>1</sup> The sense of correlation relevant here is any consistent correspondence between values of the input and output. Unlike the usual product-moment correlation or even nonparametric correlation measures, metrical or ordinal values of variables need not be preserved – only correspondences. This sense of information is one that is captured by the shared information or shared entropy measure  $H(x,y)$  discussed in (Attneave, 1959). I should note here that the requirement of carrying information is a necessary but not sufficient condition for explicating the mind-world relation. A great deal more is needed. In particular., while the information measure may not be sensitive to consistency of correspondence or even preservation of relative magnitudes, other considerations may further constrain the nature of the correspondence mapping (for many purposes, for example, the mapping has to be at least homeomorphic or local-neighborhood-preserving).

at least clarification) is central to our understanding of how the mind connects with world in perception.

It is now widely accepted in Cognitive Science (as well as in Computer Science) that many generalizations cannot be stated without recourse to the notion of representational content: Many of the things we do can only be explained if we refer to how we represent the world, what we see it *as*, what beliefs and goals we have. There is of course, much to argue about here (especially if you are a philosopher), but it will scarcely come as a surprise to a cognitive scientist to be told that, for example, the reason you where you are at this particular room at this particular time is because of what you believe and what your current goals are. Even without appealing to such notions as beliefs and goals, vision science has to refer to perceptual contents. As the examples provided by Julian Hochberg (Hochberg, 1968) nicely illustrate, how you see a certain part of a scene (what you see it *as*) depends on how you see some other part of the scene. How you see one particular line in a drawing determines (or at least constrains) how you see another line. What color you see this patch of a stimulus to be affects what color you see this other patch to be, regardless of the physical causes of the color perceptions. Many perceptual regularities have to be stated over how things appear to you; in other words, over how things are represented.

The need to appeal to representational content results in another explanatory puzzle, beyond the one that led to Kepler's problem of the inverted image. Not only do we need an informational view of sensory encoding, we also need a way to talk about representational content. A complete story of perception ought to have something to say about why some perceptual state is about X (has the content X) as opposed to being about Y. For Hume (and presumably for Kepler) what makes an internal state a representation of X is that it *looks like X*. But if "looks like" is replaced by "carries information about" then the problem of where the content comes from must be confronted once again because the information-carrying relation is concerned only with correlation and there are an indefinite number of properties of the world that are correlated with the internal states of the mind (e.g., the temperature, or most other properties of the objects). Picking out the right one is hard enough, but the content is not only some thing in the world, it is also the way the thing is seen (e.g., the Necker cube pattern of Figure 4-1 is correlated with two different percepts, so how do we specify the different percepts in terms of properties of the world?). To vision scientists who take representations and representational content for granted this question generally does not arise. The implicit understanding is that what representations represent is in some way traceable to what caused them, or at least what might have caused them in a typical setting (the latter qualification is also understood because without it we would be hard put to explain illusions or representations of imagined things that do not originate from immediate causal links with the perceived world). While this is certainly a reasonable starting assumption it is incomplete in crucial ways since there are generally very many ways that any particular representation could have been caused, yet the representation may nonetheless unambiguously represent just one scene.

Although it may seem at first glance that we should be able to give a purely mathematical account of what all these causal antecedents have in common (for example we should be able to provide a geometrical account of all the distal objects that result in a particular representation or appearance) this turns out not to be the case because the mapping between distal shape and proximal image (or, perhaps more perspicuously, between proximal information and percept) is indeterminate, or the mapping is not reversible. What something looks like (even if we could state that with unambiguous precision) depends on factors other than the geometry of the proximal image. In recent years significant progress has been made in making such factors explicit, and the current state of understanding the relation between the geometry of the proximal image and the perceived 3D shape is relatively advanced (see, for example, Koenderink, 1990;

Marr, 1982), yet we are still far from having an account of why we see things the way we do, let alone why certain of our brain states are about some things and not others. Indeed it is not clear what sort of answer might be adequate for the latter, which may account for why neuroscience celebrates findings of topographical projections of a scene as among the clearest exemplars of (at least visual) representation. But the Humean idea of representational content being defined in terms of similarity will not suffice – as anyone who has taken an introductory course in philosophy of mind knows, *similarity* is the wrong sort of relation to bridge the gap between the world and its representation (many things are similar but do not refer to one another and many things, such as words, refer without bearing any similarity).

There are at least two distinct kinds of relations between mind and world. There are semantic or referential relations and there are the causal relations. The first is the sort of relation that exists between, say, a sentence and what it expresses (its content). This is sometimes referred to as the relation of satisfaction – if the sentence is true the world satisfies the sentence (or, put it the other way round, what the sentence expresses is a state of affairs that would satisfy the sentence). The second sort of relation is one that concerns the mathematician, physicist and biologist – it is the one to which Kepler contributed important insights and the one that continues to be the goal of neuroscience (at least at the present time – I know of no principled reason why the vocabulary of neuroscience cannot be broadened to encompass the first sort of relation, the relation of content). One of the perennial projects in philosophy of mind has been to reconcile these two, presumably by showing how the intensional is grounded in the causal. Despite some impressive progress I think it is fair to say that the results have been limited. One elaborate theory has been concerned with the question of how the referents of proper names are grounded in a series of causal links to an initial a dubbing or “baptism” event (Kripke, 1980). Another theory builds on the concepts of information and information-carrying states (developed by a number of people, but perhaps best represented by the work of Dretske, 1981). In contrast, the causal connection between the proximal pattern (e.g., the distribution of light on the retina) and the three-dimensional layout of the world is well enough understood in principal, although of course there is an enormously complex story that would have to be told to explain how it works in particular circumstances. This is an area of cognitive science where considerable progress has been made, on many fronts, in the past 50 years: on the optical front – including the study of the relation between 3D geometry; the material composition of surfaces and the patterns of light that they reflect to the eye; on the biological, cellular and biochemical processes that take place in the eye itself, on the psychophysical relations that hold between optical and geometrical properties and perceived properties, as well as on the neural circuits leading from the eye to the cortex via several distinct pathways (and to a lesser extent past the primary cortex to the misnamed “association area”, where a lot of computing but little or no actual association occurs). Much remains to be discovered, but at least in the short term the kind of story it will be is unlikely to rest on brand new concepts, as it did in the time of Kepler and Descartes, when some of the basic concepts we now take for granted were missing.

The semantic or *intensional*<sup>2</sup> connection is quite a different matter. Philosophers (with a few notable Platonist exceptions) have understood that when you postulate representations – as everyone in cognitive science does – you are assuming that the contents of the representation correspond, or could correspond, in some way to entities and properties in the world, or at least in some possible world. Yet there is no straightforward way that the world causes the particular contents that our representations have, at least not in any transparent way; rather the world may *satisfy* the representation, or the representation may be true of the world. A moment's reflection should convince you that if you claim to have a theory of how the world causes your representation to be about X rather than Y the account would be missing something. For one thing the very same world pattern (e.g., of a Necker Cube) can be perceived as (represented as) one sort of thing at one time and another sort of thing at another. Psychology is full of examples where what you see something as is not determined solely by how or what it is. Illusions provide convincing demonstrations of this, but the principle runs through normal veridical perception. In (Pylyshyn, 2003, chapter 1) I provide many examples of this principle, including examples from color-mixing (the “laws” of color mixing apply over perceived colors, not over spectral properties) and shape perception, that show that how one perceives one part of a scene depends on how one perceives (represents) another part. This is not the place to rehearse these examples, but it should be kept in mind that the question of *how something is represented*, or what it is represented as, is at the heart of the study of cognition – one might even say that it is constitutive of cognitive functioning. Examples are not hard to find: It was not the holy grail that caused the knights of the round table to go out on their searches, but rather the knights' beliefs about the grail, and those beliefs have no causal connection with the grail (since there is no grail to be causally connected to). The need for talk about representations is completely general and unavoidable in cognitive science (see, for example, the discussion in Pylyshyn, 1984). Because of this it has often been assumed (and at one time it was argued explicitly by Fodor, 1980a) that an account of cognitive processes begins and ends with representations. The only exception to this, it was assumed by many (including, implicitly, in Pylyshyn, 1984), occurs in what are called transducers (or, in the biological literature, “sensors”), whose job is to convert patterns of physical energy into states of the brain that constitute the encodings of the incoming information. According to the computational view of mind, which these days represents the most widely accepted foundation of cognitive science, these states enter into the causal story of how the brain computes – how it makes inferences and decisions and ultimately determines behavior. Given the view that the bridge from world to mind resides in transduction, the problem then becomes to account for how transduced properties become representations, or semantically evaluable states and, in particular, how they come to have the particular representational content that they have;

---

<sup>2</sup> The terms “intenSional” and “intenTional” are used in a somewhat special sense. *Intensional* (with an S) is a term that appears in discussions of semantics and is used to refer to the meaning (or sense) of an expression as opposed to the things in the world of which it is true (its *extension*) – so it is relevant to the *form* of a mental representation. On the other hand *Intentional* (with a T), while often used informally to mean “done with some particular intent” is also a technical term introduced by Brentano (and discussed by Sartre and Husserl), which refers to the property that mental states and acts have in virtue of which they directed at or are *about* something (see next paragraph).

how, for example, when confronted with a red fire engine the transducers of the visual system generate a state that corresponds to the percept of a red fire engine and not a green bus.<sup>3</sup>

The problem arises because of the way that representations are related to what they represent – to how their contents are related to the world. Representational content is related to the world semantically, by the relation of *satisfaction* and satisfying is very different from causing. Satisfaction is the relation that holds between a description and the situation being described. Franz Brentano (Brentano, 1995 /1874) understood that this sort of relation is unique to the study of mind; it does not appear in physics, chemistry or biology. Because of this it presents special problems for the scientist – problems that are unappreciated by many people working in empirical cognitive science where it has typically been assumed that the causal story, or at least some abstraction over the causal story, will eventually render obsolete such distinctions as those between satisfying and causing. But the question of how the semantic relation can be naturalized remains as deep a mystery as we have in the field.

Needless to say, I will not be taking on what Brentano called the problem of intentionality. I will instead confine myself to a very small corner of this problem. Yet it is a corner that has wide ramifications throughout cognitive science. In trying to make headway in understanding the distinction between the causal and the semantic connections – between causing and satisfying – I will draw heavily on empirical findings as well as ideas from computational vision. Many of these results come from over three decades of experimental research in my laboratory as well as my earlier attempts to build computational models with computer science colleagues. Others come from recent experiments by psychophysicists and cognitive neuroscientists around the world.

What this work highlighted for me is that at the core of the connection between mind and world lies the question of how vision is able to select or pick out or refer to individual *things* in a scene – tokens or individuals rather than types. It turns out that on this seemingly simple problem rest many deep issues, from the set of problems concerned with re-identifying individual things in the world, often referred to collectively as the correspondence problem, to the grounding of concepts in nonconceptual relations to the world, and perhaps even the problem of sentience itself. (This may be a good place to interject a note about terminology. I often use the term “things” because that makes it clear that I am not intending a technical term, but at other times, when I want to invoke the usage in Philosophy or psychology, I may call them sensory individuals or visual objects or sometimes just objects. The question of what these things really are is obviously of central concern and will be addressed in due course.)

---

<sup>3</sup> At one time it was seriously contemplated that this was because we had a red-fire-engine transducer that caused the red-fire-engine cell to fire which explained why that cell corresponded to the content “red-fire-engine”. This clearly will not work for very many reasons, one of which is that once you have the capacity for detecting red, green, pink, ... and fire-engines, buses, and so on, you have the capacity to detect an unbounded number of things, including green fire-engines, pink buses,... In other words, if you are not careful you will find yourself having to posit an unlimited number of transducer types because without serious constraints transduction becomes productive. Yet even *with* serious constraints on transduction (such as proposed in Pylyshyn, 1984, chapter 9) the problem of content still remains. How do we know that the fire-engine transducer is not actually responding to wheels or trucks or engines or ladders, and so on. Any of which would do the job for any finite set of fire-engines. This problem is intimately tied up with the productivity and systematicity of perception and representation. Failure to recognize this is responsible for many dead-end approaches to psychological theorizing (Fodor & Pylyshyn, 1981; Fodor & Pylyshyn, 1988). xxx

What I hope to do in this introductory chapter is introduce this family of issues in two ways. First I will recount an early experience I had in trying to build a computer system that could reason about geometry by drawing a diagram and in the process notice particular properties of what it was drawing that could lead to conjectures about more general necessary properties and thus to possible lemmas to prove. I confess that we did not get very far along that particular road, but thinking about this problem did serve to alert us to some of the prerequisites for making progress and it is these prerequisites that I want to share with you. After this introductory example I will outline a number of apparently diverse phenomena in vision that raise the same problem – the need for a nonconceptual connection between thoughts and things in the world. Following this I will sketch the theoretical idea of a mechanism within the visual system that I call a visual index or FINST that arose from this experience, and I will describe some experiments involving attentional selection and multiple object tracking that illustrates the function of this mechanism fairly directly. In subsequent chapters I will expand on the points raised here and develop them in a way that makes contact with some contemporary philosophical issues. In every case, however, I will keep close to the empirical phenomena that motivated the initial exploration of these issues.

### **1.3 The need for a direct way of referring to certain individual tokens in a scene**

#### **1.3.1 Incremental construction of representations (and a brief sketch of FINSTs)**

Many years ago I was interested in the question of how diagrams function in reasoning. So, together with my computer science colleague, Edward Elcock, we set ourselves the ambitious goal of developing a computer system that would conjecture lemmas and prove theorems in plane geometry by drawing a diagram and noticing interesting adventitious properties in the diagram (this work was described in Pylyshyn, Elcock, Marmor, & Sander, 1978). Since we wanted the system to be as psychologically realistic as possible we did not want all aspects of the diagram to be “in its head” but, as in real geometry problem-solving, remain on the diagram it was drawing and examining. We also did not want to assume that all properties of the entire diagram were available at once, but that they had to be *noticed* over time as the diagram was being drawn and examined. If the diagram were being inspected by moving the eyes, then the properties should be within the scope of the moving fovea. Even without the eye movement complication, what is noticed has to be constrained in some way so that some degree of sequential construction of a representation is necessary. Consider the following problem that these constraints immediately raised.

Suppose the system began by drawing a line, then another line, then a line that happens to intersect a line that was already there, forming a vertex, illustrated in Figure 1-1.

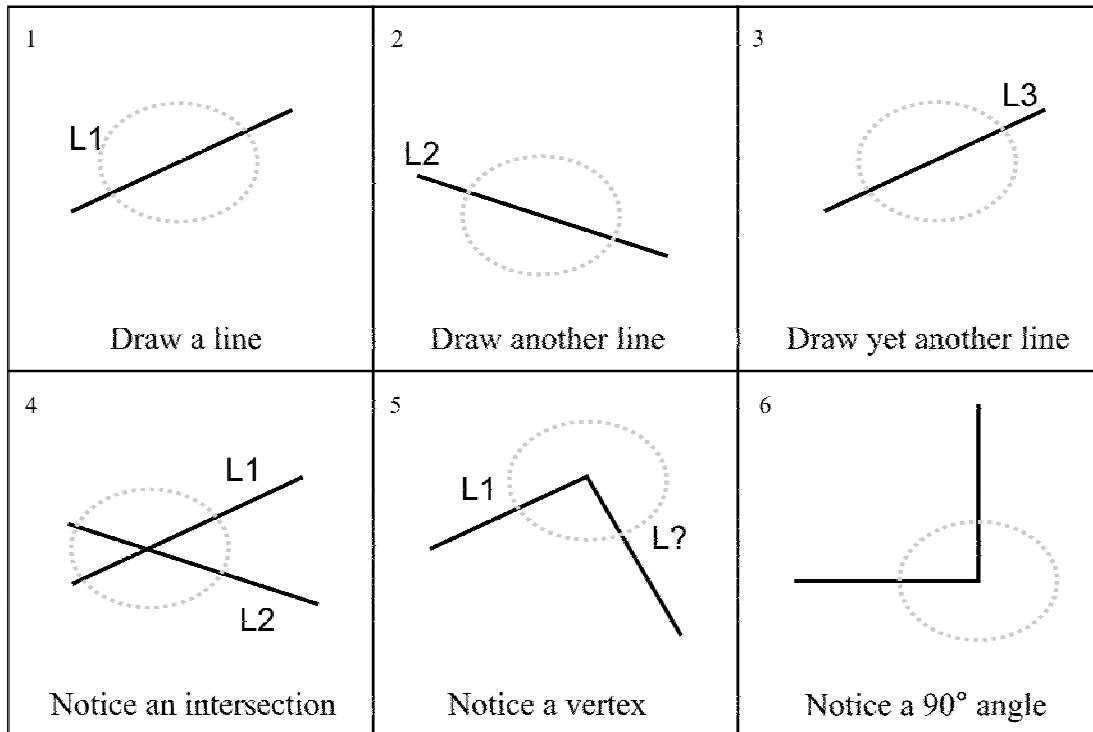


Figure 0-1. As we draw lines (which we see through a narrow foveal view shown by the ellipses) we need a way to refer to particular ones. We can do that by associating them with a description (e.g., "... is at 28° from horizontal") or by placing a label near them. Now what else do we need to re-recognize them when they recur as an intersection or a vertex, or when a second vertex is recognized, or when another property of a vertex (e.g., being 90°) is noticed?

Assume that as these three lines and the first intersection were drawn, representations of them were constructed in working memory (the memory where active representations are stored while they are being used). Working memory now contains a representation of three lines and a vertex. But do we know which line is which, and which of the represented lines form part of the vertex? Since we have drawn three lines at this point we can infer that the vertex involves two of these lines, but which ones? And of the two that form the vertex, which is which? So far it hardly seems to matter. We can easily distinguish them by their orientation. But what if we could not – what if two of them had the same orientation (as in the first and third line in this example)? Surely we know that there are two lines and that one was drawn before the other, but how do we represent this fact? We might recall where the lines were in some global (allocentric) frame of reference. But there is reason to think that we cannot localize things in a featureless global environment very well. And even if we could, knowing their location would not help if they were moving around (a common condition we will explore later). In general what we need is to be able to refer uniquely to the individual lines so as to think "*this* line was drawn first". In other words we need a way to refer uniquely to a token item (line, vertex, endpoint, etc) in the scene.

To pursue this story, suppose that the system scans the figure being drawn and notices a vertex that looks to be a right angle (as in Panel 6). Is this the same vertex as was just examined or is it another vertex that was not seen before, or which may have been noticed before but not encoded as a right-angled vertex? As the figure grows in complexity the question of whether some newly-noticed property is a property of a new or a previously noticed thing, becomes more

difficult to decide and the number and precision of properties that we would have to store in order to tell which line or vertex was which would have to grow. In order to tell, say, that the line labeled L1 in the first panel of Figure 1-1 is a different line from the line labeled L3 in the third panel, but the same as the line we have conveniently labeled as L1 in the fourth panel, we would need to encode it as a line and then check that line against each line encountered so far and determine whether it is that one by referring to its defining properties (e.g., its orientation or its location in the scene). We will see later that there is very good empirical evidence that under many common circumstances we do not re-recognize a token thing as the same identical thing previously encountered by checking its properties, and that indeed we *could not* in general do it this way because of the intractability of the problem of storing unique descriptions and matching such descriptions to solve the identity problem (or as it is known in vision science, the “correspondence problem”). Moreover the properties of items often must be ignored, as when we notice only the configurational pattern that holds among tokens and not the properties of individual tokens (in determining, for example, whether there are things in a display that are collinear).

But the situation is even worse than this characterization suggests because the same questions arises in the case of things whose properties change over time. The world is dynamic and some individual thing you now see that has a certain shape and color and is at a certain location may be the very same thing that you later see with a different shape, color, or location. It turns out that this problem is completely general since the same individual can look different at different instants in time and will clearly be in different locations on the retina and perhaps in the world. The problem I have just hinted at arises from the fact that standard forms of representation can only refer to a *token individual* by picking it out in terms of a description that uniquely applies to it. But how do we know which description uniquely applies to a particular individual and, more importantly, how do we know which description will be unique at some time in the future when we will need to find the representation of that particular individual token again in order to add some newly-noticed information to it? Moreover when we need to determine whether some particular object is the same as one seen earlier we would have to look it up by the description it had been given earlier – but how do we know what description it had been given earlier that made it unique in that previous context?

This problem of keeping track of individual token things by using a record of their properties is in general intractable when the things can move and change properties. But the problem exists even for a static scene since our eyes are continuously moving and the lighting changes with different points of view and so on – which means that the problem of unique descriptors applies to every thing in a perceived scene. In fact it remains even if the scene and the point of view are fixed (as when a static scene is viewed through a peephole) since the representation itself is changing over time as the scene is explored with moving focal attention. There is ample evidence that percepts are built up over time. It takes time for certain perceptual phenomena to appear; for example, it takes up to half a second for hidden parts of simple figures to be filled in or for illusions to be produced from the context of the figure (Reynolds, 1981; Reynolds, 1978; Schulz, 1991; Sekuler & Palmer, 1992), as well as for visual processes such as those involved in the perception of faces (Bachmann, 1989; Calis, Sterenborg, & Maarse, 1984; Hagenzieker, van der Heijden, & Hagenaar, 1990) to complete. All these phenomena require that tokens of individuals – parts of figures or other token things – be tracked so that the information developed over time can be properly merged and attributed to the appropriate things in a scene.

For now my argument concerns the sort of re-identification or correspondence computed by the visual system in the course of normal perception of scenes over relatively brief times. It does not apply when you recognize objects after some absence, as when you recognize someone you



have not seen for some time. There are clearly many cases where re-recognition proceeds by matching information stored in long-term memory and in which re-recognition fails when properties of the individual change. The present discussion concerns the sort of tracking of identity that occurs automatically and generally unconsciously as you perceive a scene and while scanning it with your gaze or your attention. It is a function of what we call early vision (Marr, 1982) or of the modular visual system (Pylyshyn, 1999). When we look at some empirical examples in the next chapter we will see the sort of time scales and conditions over which this operates.

When we first came across this problem in the context of incrementally constructing a representation of a geometrical diagram it seemed to us that what we needed is something like an elastic finger: A finger that could be placed on salient things in a scene so we could keep track of them as being *the same token individuals* while we constructed the representation, including when we moved the direction of gaze or the focus of attention. What came to mind is a comic strip I enjoyed when I was a young comic book enthusiast, called Plastic Man. It seemed to me that the superhero in this strip had what we needed to solve our identity-tracking or reidentification problem. Plastic Man would have been able to place a finger on each of the salient objects in the figure. Then no matter where he focused his attention he would have a way to refer to the individual parts of the diagram so long as he had one of his fingers on it. Even if we assume that he could not detect any information with his finger tips, Plastic Man would still be able to think “this finger” and “that finger” and thus he might be able to refer to individual things that his fingers were touching. This is where the playful notion of FINgers of INSTantiation came on the scene and the term FINST seems to have stuck.

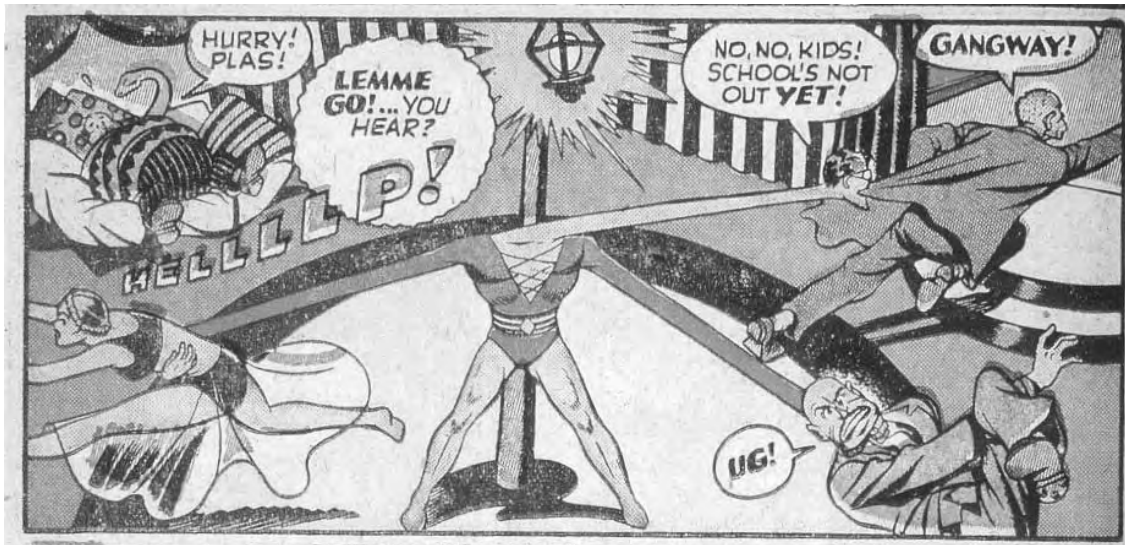


Figure 0-2. *Plastic Man* is able to extend his limbs flexibly. Even if his tactile sense did not permit him to recognize what he was touching, he would still be able to keep track of things in the world as the same individual things despite changes in their location or any of their perceptual properties. (From “Police Comics” #21 © DC Comics. PLASTIC MAN™ DC Comics. All rights reserved. Used with permission)

### 1.3.2 Using descriptions to pick out individuals

I have been speaking of the need to keep track of things without using their properties, or more precisely, without using a conceptual description. But how can we keep track of a thing unless we know something about it? In particular, how can we keep track of it unless we know *where* it is? What I will suggest in the next chapter is that *selection*, which is the central

function of what has always been called *focal attention*, is based on individuals, which in vision means that it is “object based” or sensitive to the individual token and not to its properties. But for now let us reconsider the geometry example and ask how we might attempt to keep track of individual parts of the figure by using a stored description. This requires that we be a bit more precise about what constitutes a description. The everyday sense of a description is both too strong and too weak. It is too strong for our purposes because it implies that there is a description in some natural language, whereas we do not need that restriction in the case of a mental representation. All we need is that a description be constructible from basic concepts (other restrictions, such as compositionality, are also required but will not be discussed here – see, e.g., Fodor, 1998; Fodor & Pylyshyn, 1988). So for our purpose a description is any encoded representation that picks out some individual token by referring to properties that it possesses. The question is: Can such a description uniquely pick out and refer to a token individual under a wide range of circumstances – in particular can it refer to an individual token under conditions such as those that we were concerned with in the geometry example? Even if it can, a second question is: Is this how the visual system does it?

In the example sketched earlier, where we are constructing a description of a figure over time, we need to keep track of individual things so as to be able to determine which is which over time – i.e., we need to be able to decide between “there it is again” and “here is a new one”. We must be able to do this in order to put new information into correspondence with the right individuals already stored in memory. We also need to be able to decide when we have noticed a new individual thing or merely re-noticed one we had already encoded earlier. Being able to place individual things into correspondence over time – or to keep track of individual tokens – is essential in constructing a coherent representation. When we notice an individual thing with property P we must attribute P to the existing representation of that very token (if we had encoded it before), or else we must augment our stored representation to include a new individual thing. One way to place individual things into correspondence is to associate a particular token thing with what Bertrand Russell called a definite description, such as “the object x that has property P” where P uniquely picks out a particular thing. In that case, in order to add new information, such as that this particular thing also has property Q one would add the new predicate Q to the representation of that very thing<sup>4</sup>. This way of adding information would require adding a new predicate Q to the representation of an object that is *picked out by a certain descriptor*. To do that would require first recalling the description under which x was last encoded and then conjoining to it the new descriptor. Each time an object was encountered once again, we would somehow have to find the description under which that same object had been encoded earlier.

The alternative to this unwieldy method is to allow the descriptive apparatus to make use of the equivalent of singular terms or names or demonstratives. If we do that, then adding new information would amount to adding the predicate  $Q(a)$  to the representation of a particular thing  $a$ , and so on for each newly noticed property of  $a$ . Empirical evidence that I will review below

---

<sup>4</sup> The way a mechanism based on updating descriptions would be used to solve the correspondence problem would be something like this (using the predicate calculus notation is inessential but convenient). The perceptual system notices an individual with property P and stores a description that identifies it at that time, say  $\exists(x)P(x)$ . If the description is to specify that P picks out a *unique* object it would have to be augmented to  $\exists x\{P(x)\wedge[\forall(y)(P(y)\supset(x=y))]\}$ . When an additional predicate Q that pertains to the same object is to be added, the previously stored descriptor for that object is retrieved and a new expression added that asserts that the object also has property Q, thus:  $\exists x\exists y[P(x)\wedge Q(y)\wedge x=y]$ . The process is repeated each time a new property of some object is encountered. The process of continual inferring the previous unique description and updating it with the currently noticed predicate is clearly not a plausible way to incrementally build a visual representation. Even if there was some rational way to determine the previous description, the matching and updating process demands increasingly complex storage and retrieval processes based on pattern matching, a process that is in general computationally intractable for tree-matching structures.

suggests that the visual system's Q-detector recognizes instances of the property Q *as a property of a particular visible object*, such as object *a*. This is the most natural way to view the introduction of new visual properties by the sensorium. This view is consonant with considerable evidence that has been marshaled in favor of what is referred to as "object based" attention and I will have more to say about this idea in the next chapter. In order to introduce new properties in that way, however, there would have to be a non-descriptive way of picking out *a*. This is, in effect, what the labels on objects in a diagram are for and what demonstrative terms like "this" or "that" allow one to do in natural language so what I am in effect proposing is that the visual system needs such a mechanism of demonstratives.<sup>5</sup>

The object-based view of how properties of objects are detected and encoded would suggest that when we detect a new property we detect it as applying to a *particular* object, rather than as applying to any object that has a certain (recalled) property. It is also more plausible that properties are detected as applying to particular objects since it is objects, rather than empty locations, that are carriers of properties – as I will argue in the next chapter. Intuitions, however, are notoriously unreliable so later I will examine empirical evidence that this view is indeed more likely to be the correct one. For example, in Chapter 2 I will describe studies involving multiple-object tracking that make it very unlikely that objects are tracked by regularly updating a description that uniquely picks out the objects.

The empirical part of this story is the hypothesis that what perception initially detects is *things* or *objects*, as opposed to properties or locations. The more general claim that *something* in the world is detected without prior specification of its properties is more than an empirical hypothesis. In order to be able to provide an explanation of behavior and its relation to environmental conditions we must allow for a purely causal connection from world to mind. Later we will see that in principal there are two ways in which properties of the world may affect a perceptual system. It may affect it in a purely causal manner. A property P in the world can simply trigger a chain of events that culminates in some change in the perceptual system. Alternatively, the perceptual system may, in effect, *ask* whether property P is present. The first of these corresponds to what in computer systems is called an *interrupt* while the second corresponds to a *test* for P. We often refer to the first as bottom-up and the second as top-down. What is important for us is that there is no such thing as a purely top-down process, or rather, a process cannot be top-down all the way out to the world. If representations are to have a content that is about the world, then the world must impose itself upon the perceptual system – which is to say it must act bottom up at some stage. What I am proposing here is that what is bottom up is what will be needed to produce the predicate-argument pairs that constitute a conceptual encoding of the world (encoding that something has the property P). In order to prevent circularity the arguments of such predicates must be identified (or as I say "picked out") by a process which itself is not conceptual (does not use other predicates or properties in order to identify the referents of the arguments). This desideratum also entails that things that are bearers of properties must be selected and referred to in a bottom-up or data-driven manner. I will return to this topic in the next chapter where I hope to show that this requirement is totally reasonable and in a certain sense even obvious.

---

<sup>5</sup> Christopher Peacocke has pointed out to me that both *demonstrative* and *name* are misleading ways of referring to Indexes. Typically a demonstrative term is voluntarily assigned and it carries the implication that what it refers to depends on the intention of the speaker and the context of utterance, which is not the case with FINST indexes. On the other hand *name* is misleading because names allow us to think about things in their absence, whereas FINST indexes have a restricted existence, corresponding roughly to when their referents are seen (and perhaps a bit longer, because of inertial persistence of sensors). Since all such analogies are misleading I will simply refer to FINSTs or visual indexes.

### 1.3.3 The need for demonstrative reference in perception

The sort of “link” I have been referring to is very close to what philosophers have called an *indexical*. Indexicals are terms that only refer in particular contexts of an utterance. They also occur in thoughts where mental indexicals refer in the context of particular token thoughts. In natural language indexicals are instantiated by such terms as pronouns (*me, you*), temporal and spatial locatives (*now, then, here, there*), and, of particular interest to us here, demonstratives (*this, that*) which pick out particular token individuals. Since my concern will be only with the selection of things, and not with other sorts of indexicals I will follow common practice and use the term *demonstrative* rather than indexical.

The easiest way to see what this sort of link is like is like is to think of demonstratives in natural language – associated with words like *this* or *that*. Such words allow us to refer to things without specifying what they are or what properties they have. While this gives a flavor of the type of connection we will be discussing, equating this sort of reference link with the role of certain words in a natural language is misleading in many ways. What a word such as “this” refers to in discourse depends on the intentions and state of knowledge of the speaker (as well as the speaker’s beliefs about the state of knowledge of the hearer). Such terms typically occur together with nouns, so we speak of “this chair” or “that table” and so on, and in such contexts they can pick out extremely general things that include things not in our perceptual field, as when we say “this house” while pointing at a wall or “this city” while pointing out the window. Such complex demonstratives occur frequently and there is even a lively debate about whether all uses of demonstratives involve (unstated) complex demonstratives or whether there can be “bare demonstratives” (e.g., Lepore & Ludwig, 2000). We need not enter this particular debate since what I am proposing is clearly not identical to a demonstrative in a natural language. To the extent that it is like a demonstrative it is clearly like a “bare” demonstrative – it picks out things without doing so by their properties. It does it because the perceptual system is so constituted that things of certain kinds and not other kinds are picked out in certain contexts. Spelling this out will be left for a later chapter but the details clearly rest on empirical findings concerning such questions as how attention is allocated and how the world is parsed and indexed.

The study of the connection between demonstrative thoughts and perception has been a central concern in philosophy of mind. Most philosophers acknowledge that demonstrative thoughts are special and essential to linking mind and world. They also recognize the important role that perception plays in establishing such links – through what are referred to as “informational links”. Many philosophers have also argued that in order to link perceptual representations to actions, individual things in a scene must be selected and that such selection requires demonstrative reference. A reason given is that, finally, the motor system must act on things that are picked out directly rather than by description. We are able to reach for ...*that*.. without regard for what that is. We can reach for it without representing any its properties since those are irrelevant to reaching for *it*. Of course the motor system must issue commands in some quantitative frame of reference, but as we will see in Chapter 5, this need not be in a global frame of reference nor in any frame of reference available to other parts of the nervous system. How the visual system can provide the information to command an eye or limb movement when the mind does not know where the item is located is a puzzle that is more apparent than real, as we will see later.

John Perry (1979) has argued that such demonstratives are essential in thoughts that occasion action. Perry offers the following picturesque example,

The author of the book *Hiker’s Guide to the Desolation Wilderness* stands in the wilderness beside Gilmore Lake, looking at the Mt. Tallac trail as it leaves the lake

and climbs the mountain. He desires to leave the wilderness. He believes that the best way out from Gilmore Lake is to follow the Mt. Tallac trail up the mountain ... But he doesn't move. He is lost. He is not sure whether he is standing beside Gilmore Lake, looking at Mt. Tallac, or beside Clyde Lake, looking at the Maggie peaks. Then he begins to move along the Mt. Tallac trail. If asked, he would have to explain the crucial change in his beliefs in this way: "I came to believe that *this* is the Mt. Tallac trail and *that* is Gilmore Lake". (Perry, 1979, p4)

This point is important and easy to overlook. In fact it was glossed over in the earlier discussion of the need to keep track of individual visual objects, illustrated in Figure 1-1. There I labeled the vertices and lines and suggested that what we needed in order to encode the diagrams over time in a coherent manner is what such labels provide. While labels help in thought and in communication, they can do so if (and only if) we have an independent way to refer to the things to which the labels apply. As in Perry's example, we can think about the labeled items if we can think thoughts such as "*this* is the line labeled L1". If we cannot refer to the line in our thoughts independently of their printed label then we cannot use the information that the label provides! Even being able to think of a line as "the line closest to Label L1" will not do because determining which line is closest to the label requires referring to the line in question directly, as in "*this* is the line closest to Label L1". The alternative would be to search for something that is a line and that is closer to L1 than any other line. But that too requires having in mind the thoughts "*this*<sub>1</sub> line is x distance from label L1 and *this*<sub>2</sub> line is y distance from L1..." We may have no awareness of such thoughts, but unless we can entertain thoughts with such contents (however expressed) we could not make use of the labels. The importance of demonstrative identification has been recognized for some time and it has also been the source of humor in such cartoons as the person lost in the desert who comes up to a sign with an X on it and the words "You are here," or one of my favorites (by David Sipress, which appeared in the New Yorker on April 22, 2002) shown in Figure 1-3. The problem illustrated in these examples is generalized in the next section to the claim that we need a way to bind representations of individual things to the token things themselves – we need a symbol-to-world binding mechanism.



*"Are we in this Starbucks or the one down the street?"*

Figure 1-3 © The New Yorker Collection 2002 David Sipress, from cartoonbank.com. All Rights Reserved. New Yorker, Used with permission

The point of this discussion is that the mental representation of a visual scene must contain something more than descriptive or pictorial information in order to allow re-identification of particular individual visual elements. It must provide what natural language provides when it uses names (or labels) that uniquely pick out particular individuals, or when it embraces demonstrative terms like “this” or “that” (though see Note 6). Such terms are used to indicate particular individuals. Being able to use such terms assumes that we have a way to *individuate*<sup>6</sup> and keep track of particular individuals in a scene *qua* individuals – *i.e.*, even when the individuals change their properties, including their locations. Thus what we need are two functions that are central to our concern in this book: (a) we need to be able to pick out or individuate distinct individuals (following current practice, when discussing the experiments I will call these individuals *visual objects*, reserving the more general question of what they really are for the later discussion) and (b) we need to be able to *refer to* these visual objects as though they had names or distinct demonstratives (such as *this*<sub>1</sub>, *this*<sub>2</sub>, and so on). Both these purposes are served by the proposed primitive mechanism that I have called a *visual index* (or more generally a *perceptual index*) or a FINST.

I might point out that even though the postulation of FINST indexes arose from the theoretical need for something like demonstrative indexes to fill the gap between symbolic representations (perhaps in the Language of Thought) and perceived physical things in the world, such indexes can now be seen as important for explaining certain human skills. Among these is the ability to play team sports, like basketball or hockey, in which a player must keep track of moving objects (e.g., players) as they weave around the field or ice. There have been reports of exceptional tracking ability among these “experts” and we now have evidence that tracking can be improved substantially with practice. Moreover we have evidence that people are able to track moving things far into their periphery, as long as they do not get too close to one another. Some of these results will be presented when we talk about our multiple object tracking experiments in Chapter 2. But for now it might help the reader appreciate the generality of the mechanism that I will be discussing if you think of them as connecting familiar moving things in the world with mental things, as illustrated in Figure 1-4 for the case of the game of basketball.

---

<sup>6</sup> As with a number of terms used in the context of perception research (such as the term “object”), the notion of *individuating* has a narrower meaning here than in the more general context where it refers not only to separating a part of the visual world from the rest of the clutter (which is what we mean by *individuate* here), but also providing identity criteria for recognition instances of that individual. As is the case with *objecthood* and other such notions, we are here referring primarily to perceptually primitive cases, such as ones provided directly by mechanisms in the early vision system (in the sense of the term “early vision” used, for example in Pylyshyn, 1999) and not constructed from perceptual/conceptual resources.

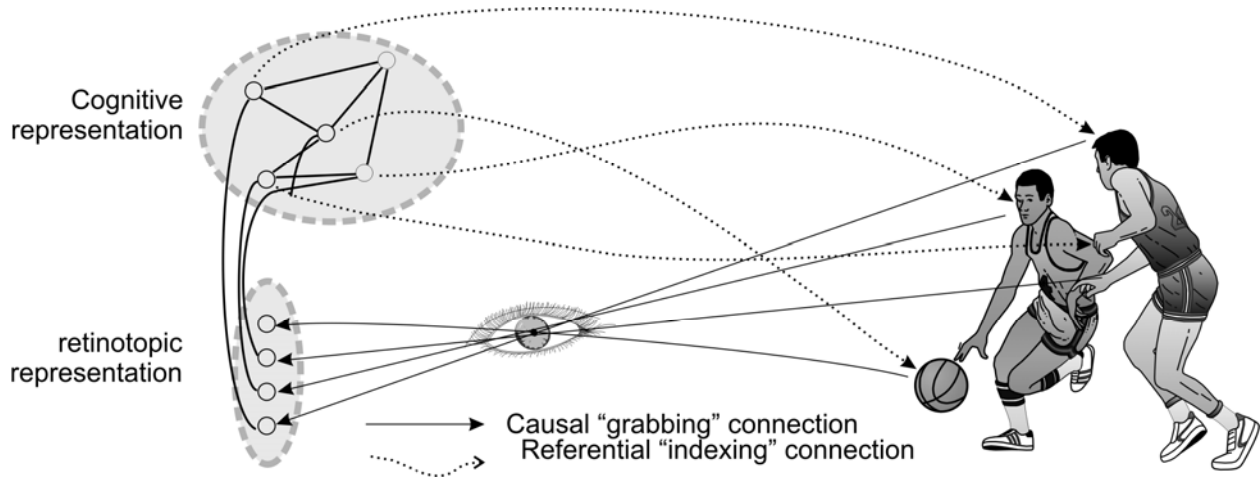


Figure 0-3. An illustration of how being able to pick out an track several individual moving things, even when there is not enough time to encode their properties, might be useful in team sports.

In the rest of this chapter I will provide some empirical illustrations of the claim that the visual system does in fact embody a primitive mechanism of this sort by showing that they provide a natural account of a number of empirical phenomena. In the next chapter I will introduce other experiments and will discuss the philosophical issues raised by this claim.

## 1.4 Some empirical phenomena illustrating the role of Indexes

### 1.4.1 Tagging/marking individual objects for attentional priority

There are a number of other reasons why the visual system needs to be able to pick out particular individuals in roughly the way singular terms or demonstratives do (i.e., without reference to their properties). This need is quite general and arises from the fact that properties are predicated of things, and relational properties (like the property of being “collinear”) are predicated of several things. So there must be a way, independent of the process of judging which property obtains, of specifying which objects the property will be predicated of. Ullman, as well as a large number of other investigators (Ballard, Hayhoe, Pook, & Rao, 1997; Watson & Humphreys, 1997; Yantis & Jones, 1991) talk of the objects in question as being “tagged” (or in some cases “marked”). One of the earliest uses of the notion of tagging was associated with explaining why things that had attracted attention (e.g., by being flashed or by suddenly appearing in the field of view) had priority in such attention-demanding processes as detecting a faint dot or making a visual discrimination. For example, (Yantis & Johnson, 1990) showed that in a search task, finding specified letters in a multi-letter display showed superior performance when the letter had been signaled (highlighted) and he attributed this to a “priority tagging” process. Tagging has also been used to explain why certain items have a *low* priority in search. Under certain conditions, irrelevant but potentially confusable distractor items can be inhibited in a search task by being tagged (Watson & Humphreys, 1997, refer to this as “marking” rather than “tagging” but the idea is the same). The notion of a tag is an intuitive one since it suggests a way of *marking objects* for reference purposes. But the operation of tagging only makes sense if there is something *out there* on which a real tag can be placed. It does no good to tag an internal representation since the object one wishes to examine is in the world (recall that one of the reasons for tagging objects is to be able to move focal attention to them, to examine them further and to evaluate predicates over them). But how do we tag parts of the world? What we need is a

way to refer to individual things in a scene *independent of their properties or their locations*.<sup>7</sup> This is precisely what FINST indexes provide.

### 1.4.2 Argument binding

When we recognize visual patterns we recognize them as patterns constituted by particular tokens. Consequently prior to detecting the pattern we must select or pick out the relevant elements of the pattern and then recognize the configuration that these elements instantiate. Shimon Ullman (Ullman, 1984) described a number of simple patterns that he claimed require, by their very nature, that a serial process (called a “visual routine”) be undertaken involving the certain selected token elements. For example in order to detect the pattern “inside” the elements which this applies to must be selected and specified. Ullman (as well as Marr, 1982) uses the notion of *tagging* to refer to this selection. Some form of a selection and specifying operation is essential because there must be some way to specify to the particular token items to which the pattern detection routine is applied. The pattern-detection process may simply involve judging whether the specified items form a particular shape (as in the left panel of Figure 1-5) or whether certain more abstract relation holds among them (as in the right panel of Figure 1-5). In the case of the more abstract relations, a visual routine such as “contour tracing” or “area painting” must then be undertaken, but this cannot be done until the things on which the process must be performed have been identified and a reference to them established. My way of putting this is to say that certain items must be bound to the argument of a visual predicate (or a computational function) before the predicate can be evaluated. In these examples we need some way to bind the arguments of predicates such as **Collinear**(x, y, z, ...) or **Inside**(x,c), as shown in Figure 1-5.

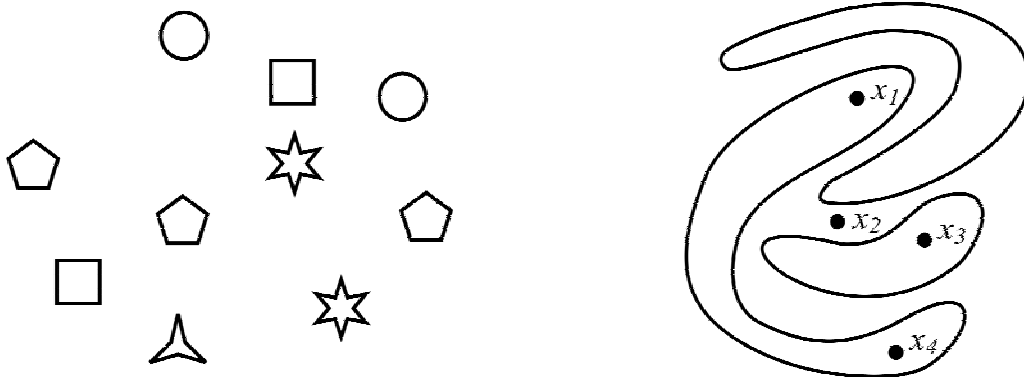


Figure 0-4, Collinearity (left panel) can only be computed over objects after they have been identified (i.e. individuated) and bound to the argument of the “collinear” predicate. Similarly the “inside” predicate (right panel) can only be computed if all relevant objects (dots  $x_1 \dots x_4$ , and appropriate contours) are bound.

In these examples the things over which the predicate is evaluated have to be selected. How does such selection occur? Is it voluntary or automatic? We will return to these questions in Chapter 2. But for the moment we might note that some form of voluntary selection must be possible. Look at a flecked wall or any surface not totally uniform. You can pick out a particular

<sup>7</sup> Actually, as we will see in the next chapter, it would not help the problem of incrementally constructing a representation if we *could* tag the objects in the world since it would not solve the problem of representing unique individuals. For example, it would not let us think thoughts such as “*this* is the object labeled L1” without which the label would be of no help. The use of demonstratives in thought is so natural that it is easy to forget that they are indispensable.



fleck or texture element with no trouble. Now pick out a second and third such fleck without moving your eyes. It is not easy, but it can be done (or, rather, it can easily be done but feels like an effort – an experience that may have little to do with how the process itself unfolds, as I will argue in Chapter 4). Experiments (e.g., Intriligator & Cavanagh, 2001) have shown that so long as the items are not too close together people can keep a particular selection and keep their eyes fixed while moving their attention to a specified second item (they can follow the instruction to “move up one” or “move right two”). We have also carried out experiments (Section 1.4.4) where the selection is automatic – where the FINST index is captured or grabbed by an onset event.

### 1.4.3 Subitizing

I want to give two additional experimental examples of the need for such argument-object binding because they make an important point about how the selection works and why it might be generally useful. Among the processes for which binding is needed is one that evaluates the cardinality of the set of tokens. There is a lot of evidence that when the number of items is 4 or less the process of recognizing their numerosity, called subitizing, involves a different mechanism from that used in estimating larger quantities. The evidence comes from both psychophysics and neuroscience and has been studied in adults, infants and animals (the latter nicely summarized in, Dehaene, 1997). While counting is involved in both the subitizing range ( $n \leq 4$ ) and the larger counting range ( $n > 4$ ), the former has certain signature properties, among which is a faster and more accurate enumeration and an independence from item location (e.g., telling the subject in advance which quadrant of the visual field the items will appear in does not alter subitizing though it does improve counting, see below). These characteristics can be explained if we assume that subitizing does not require searching a visual display for the items to be enumerated, because what is being enumerated is the number of active FINST indexes. But the explanation that involves indexes assumes that the relevant items are individuated automatically and quickly and a reference is established at the same time. When this precondition is not fulfilled subitizing cannot occur as we discovered (Trick & Pylyshyn, 1994b).

There is independent evidence that certain conditions of element properties (and spacing) allow automatic individuation and others do not. For example when items are too close together they cannot be automatically individuated but require focal attention, as evidenced by a person's inability to pick out, say, the third one from the left, even though the distances are large enough that person can easily judge when there are two items and when there is just one (the usual 2-point threshold test for acuity). See Figure 1-6.

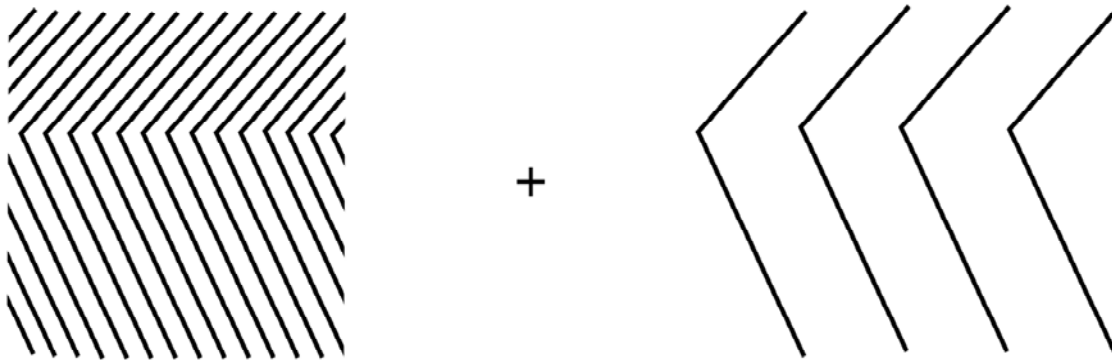


Figure 0-5. The spacing required for individuation is greater than the spacing required for two-point resolution. The first is measured by the ability to select a particular element (e.g., third from left) while the latter is measured by the ability to distinguish one from many without deployment of focal attention. (based on Intriligator & Cavanagh, 2001)

Given these independently established individuation parameters we can then ask whether elements that cannot be individuated without serially attending to them (as in the panel on the left) can be subitized. The answer we obtained from experiments is that when items are arranged so that they cannot be preattentively individuated, for any reason, they can't be subitized either, even when there were only a few of them (Trick & Pylyshyn, 1994b). For example items that are too close together or are distinguished only through operations that must be performed serially or that require serial focal attention in order to individuate (e.g., objects characterized as "lying on the same curve" or elements specified in terms of conjunctions of features, such as elements that are both red and slanted to the left) cannot be subitized. In other words with such displays we don't find the discontinuity in the rate of enumeration as the number of objects exceeds around 4 (as shown by the fact that the graph of reaction time as a function of number of items does not have a "knee").

An example of elements that can and that cannot be individuated preattentively, along with typical reaction-time curves, is shown in Figure 1-7. When the squares are arranged concentrically they cannot be subitized whereas the same squares arranged side by side can easily be subitized regardless of whether or not they are the same size. Trick & Pylyshyn argued that the difference between counting and subitizing lies in the need to search for items in the world when counting large numbers ( $n > 4$ ) of items, requires attentional scanning which takes time and memory resources. By contrast the cardinality of smaller numbers of items that have been indexed can be ascertained without having first to find them. This can be done either by counting the number of indexes deployed or by evaluating one of several cardinal predicate over them (e.g.,  $TWO(x,y)$ ,  $THREE(x,y,z)$  and so on). Since there is a (small) increase in time taken to respond correctly as number as the number increases from 2 to 4, the first of these appears more natural.

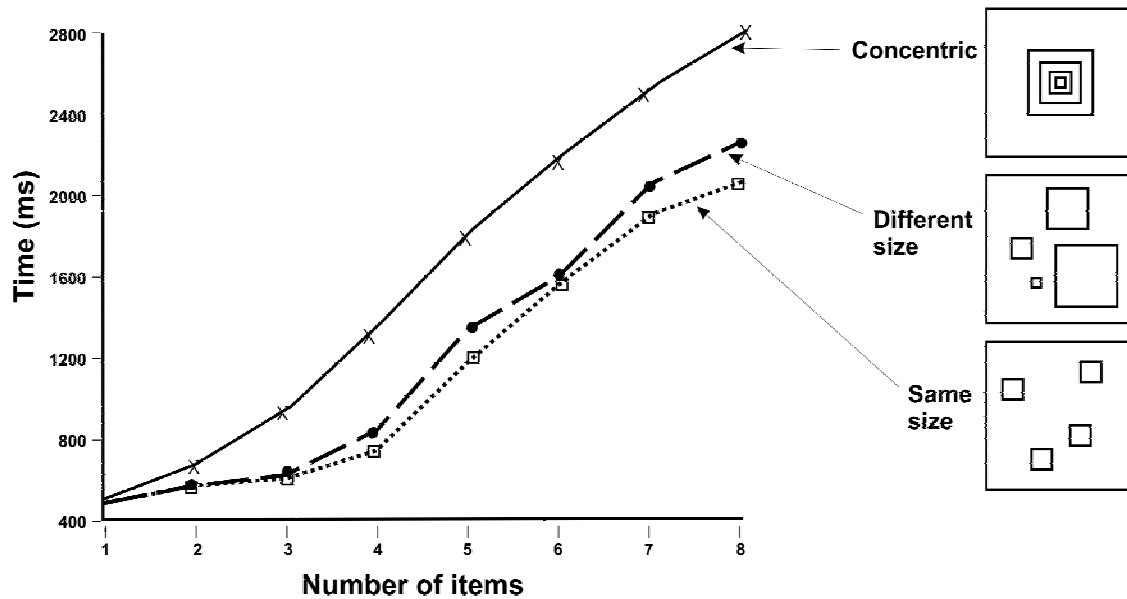


Figure 0-6. Graph of reaction time versus number of items enumerated, for several conditions examined in (Trick & Pylyshyn, 1993). Concentric squares do not show the characteristic “knee” in the curve that is the signature of subitizing.

#### 1.4.4 Subset selection

I have claimed that a central function of indexes is to select and refer to (or bind arguments to) several visual elements at once so that visual predicates can be evaluated over them. This is important not only for recognizing certain patterns as I suggested above, but if we make certain assumptions concerning how the indexing mechanism works, may also help us to understand how visual stability is attained in the face of rapid saccadic exploration of the visual world. Let me illustrate with two experiments.

The first study (Burkell & Pylyshyn, 1997) was an experiment in which a subset (of 2-5) items sprinkled randomly among a set of identical 11 items (X's) was precued (by an attention-capturing signal), following which all 11 items turned into distinct search items (by dropping one of the bars and changing colors, yielding left-oblique and right oblique bars in either red or green) and the subject had to search through only the precued subset for a specified target (e.g., a left-oblique red bar). The patterns were such that we could tell whether the subject was searching through only the precued subset or in fact ended up searching through the entire set of 11 items.<sup>8</sup> What we found is that subjects did confine their search to a subset of cued items among a larger set of similar items. Moreover, their performance in finding the target was not slowed when the distance among members of the subset was increased, as one would expect if subjects had to search through the subset items by scanning the display. These results suggest that subjects could hold the subset in mind during the search and also that they did not have to search for the subset items themselves; they only had to search for the target among those subset items. This is despite the fact that subset items were interspersed among the other (distractor) items. We concluded that the sudden onsets caused indexes to be assigned to the designated subset, which then could be used to direct a rapid search of that subset while ignoring the

<sup>8</sup> The technique involved presenting all 8 X's and then using sudden onsets of additional X's to precue a target subset of 3 to 5 Xs. All the Xs then turned into either a “popout” single-feature search or a slow “conjunction” search. Since the elements in the entire display always constituted a set of conjunction-search items, we could tell by the different search rates whether subjects were able to confine their search to the subset alone. More details on all these experiments can be found in (Pylyshyn, 2003).

irrelevant intervening items – much as the enumeration operation could be confined to the selected items in the subitizing task, providing there were 4 or fewer of them.

The second set of experiments (carried out with Christopher Currie) used the same procedure, but introduced a saccadic eye movement after the subset had been cued but before the 11 “X” items changed into search items (left- and right-leaning colored bars). In these experiments we found that under certain saccade-inducing conditions (in particular when subjects moved their gaze to any one of the targets, though not when they were told to saccade to a second fixation point or to the edge of the display) observers were still able to confine their search to the subset. This finding lends support to the proposal that what makes the world appear stable in the face of several saccades each second may be that the correspondence of a small number of items between fixations is made possible by a mechanism such as a FINST. Others have also shown that only a limited amount of information is retained across an eye movement. In fact (Irwin, 1992; Irwin & Gordon, 1998) showed that only information for about 4 objects could be retained, which fits nicely with the account we give based on FINSTs.

The power of FINSTs to select and to hold on to objects also accounts for a number of findings reported in the literature in which recalled patterns are superimposed onto patterns that subjects viewed visually. For example they can account for how illusions such as the Müller-Lyer illusion can apparently be induced by *imagining* arrows superimposed on perceived lines. All one has to assume to account for this is that indexed endpoints and arrows allow attention to be moved to them. Since the Müller-Lyer is known to depend on attention, this is enough to induce the illusion. Many such demonstrations are discussed in (Pylyshyn, 2003, chapter 5 & 7).

## 1.5 What are we to make of such empirical demonstrations?

I have devoted rather more space to these examples than perhaps is merited by the small point I wish to make. I simply want to point out that there are many reasons why the visual system needs to pick out individual token things in a perceptual scene. Moreover the picking out entails two separate operations. First, it entails a form of individuation – a primitive separation of the thing from its background and from other things. Second, it entails being able to refer<sup>9</sup> to the individual directly – in an unmediated way that does not require using a description of the thing in question. The reason for separating these two functions may not be apparent at this stage, but I will return to it in the next chapter where I distinguish them empirically, with the first

---

<sup>9</sup> John Campbell has suggested that I might avoid some philosophical arguments if I refer to FINSTs as “epistemic instruments” which serve to find out about real physical objects in the world and to act on them, rather than treat them as demonstratives or direct references, since the latter raises questions such as whether they refer to real physical objects or some “proto-objects”, whether they play an inferential role similar to proper names, and whether they are two-place relations (as implied by my term “direct reference”) or three-place relations involving a reference, an object and a some encoding of the object’s properties (e.g., an “object file”). The need for a three-place relation appears to arise because of the possibility that two distinct indexes happen to refer to the same thing so they must be individuated by other than their referents alone. These are all valid and helpful observations and I am grateful to John Campbell for his comments. For a number of reasons having to do with my expository goals (which I hope will become apparent later) I will persist in my claim that indexes directly refer to proto-objects or *things*. However the last point concerning the possibility of two indexes referring to the same thing – and the related question of how it is possible to decide whether this is indeed the case, requires some additional comment that I will take up in the next chapter. Essentially my position is that FINST indexes are distinguished by the causal history by which they come to refer, so there can indeed be multiple indexes to the same thing, and that token indexes can be distinguished, just as singular terms are – by their syntactic shape (i.e., one is  $P_i$  and another is  $P_j$ , where  $j \neq i$ ). Determining whether they refer to the same thing may or may not be something that can be done within early vision, depending, among other things, on spatio-temporal conditions. In some cases such a determination may require conceptual intervention (re-recognition may require appeal to objects’ properties).

(individuation) function being carried out in parallel and without drawing on limited resources, and the second being limited to the 4 or 5 indexes postulated in FINST theory.

## Chapter 2. Indexing and tracking individuals

### 2.1 Individuating and tracking

In the previous chapter I introduced the need for indexical or demonstrative reference to individual things. Postulating such perceptual demonstratives or indexes assumes several sorts of capacities. One is the capacity for *individuating* whatever is indexed or demonstrated. The second is the capacity for keeping track of such individuals – of tracking each of them *as being the same enduring thing* despite changes in its appearance and, in particular, despite changes in its location. These two assumptions raise both philosophical and empirical questions and even contradict a great deal of received wisdom in both fields. Consequently I need to explain my uses of terms such as individuating and tracking or re-identifying, both of which are at the heart of the present proposal.

In recent years psychologists have used the term *individuate* to indicate that some part of the world is perceptually separated from the rest by a process that is related to the Gestalt notion of *figure-ground* separation. This sort of parsing of a scene into what perception treats as objects of interest, as distinct from the rest of the undifferentiated world, is one of the most basic operations that a perceptual system performs. In fact recent evidence suggests that it is performed by the visual system of babies only a few months old (Johnson, 2001). But the exact nature of this sort of individuation is not very clear. For example, it is not clear whether in individuating something we must also notice and encode or represent particular properties of that thing. Philosophers assume that in order to individuate something we must conceptualize its relevant properties. In other words we must first represent (or cognize or conceptualize) the relevant *conditions of individuation*, otherwise how can we distinguish the individuated thing from other things in the perceptual field? The concepts that we need in order to do this, according to this story, are called *sortals*. Sortals are concepts that correspond to countable things, so they include the concepts shoe, table, chair, person, circular disk, and so on, but not water, air, sky, and other things that correspond to “mass nouns” which are not countable (you can say “some water” but not “three waters”). Carving up the world according to sortal concepts is, according to this story, a prerequisite for individuating them. Some people believe that very few sortal concepts are available in early infancy – for example it may be that babies only have one sortal concept, namely the sortal “object” (Xu, 1997). According to this view, in order to re-recognize an individual at another instant we need to be able at least to assign it to the same sortal concept.

The idea that identifying (or reidentifying) something as the same individual thing requires conceptualization was vigorously defended by the philosopher Peter Strawson (Strawson, 1959). Strawson argued that such identification requires “the apparatus of concepts” which includes not only sortal concepts, but also the conceptual apparatus of “numerical identity” needed for counting, divided reference (distinguishing *this* from *that* entails referring to at least two things at once) and tenses (in order to identify *this<sub>now</sub>* with *this<sub>before</sub>*). I agree with all of Strawson’s arguments – individuating and reidentifying in general require the heavy machinery of concepts and descriptions. What I will argue is that FINST indexes give one a special kind of approximation that serves reliably to do the work of individuation and re-identification in our sort of (restricted) world. But what FINSTs provide is not just an approximation. What they provide is indispensable for true individuation. Without a nonconceptual mechanism of the sort

provided by FINSTs we would not have the full sense of individual and re-identification. Conceptual identification ultimately needs a nonconceptual basis.

I am going to argue that, notwithstanding the claims about the need for sortal concepts, there must be a type of individuation and a type of re-identification that occurs at least in visual tracking, which is more primitive than the individuation provided by sortals. It is a nonconceptual type of individuation and tracking, and its existence is supported both by empirical arguments (from experimental evidence and from more general empirical considerations) and by philosophical considerations. I will also suggest that it is this that we see in operation both in the experimental examples I will present and also very likely in the infant studies of object-constancy and infant sensitivity to the cardinality of sets of objects (to be reviewed briefly in Section 2.5). Before getting into these philosophically-loaded questions, I will offer some empirical demonstrations of what FINSTs can do (an animated demonstration of the experimental materials can be viewed at <http://rucss.rutgers.edu/finstlab/demos.htm>).

The core idea I will explore in this chapter is the idea of *tracking* or of *keeping track of individuals*, or of recognizing what is sometimes called the *numerical identity* of a thing. Part of what it means to individuate something is to be able to keep track of its identity despite changes in its properties and location. To know that something is an individual is to know that *this* (at time  $t$ ) is the same thing is *this* (at time  $t+\Delta t$ ), so that when it changes properties or moves in certain ways over time it keeps being the same enduring individual. We know that in general this requires re-recognition. But we also know that the perceptual system does it automatically in very many circumstances without paying attention to its properties (indeed, by explicitly ignoring many of its essential properties). This happens routinely in apparent motion as well as in stereo perception, both of which require solving what has been called the *correspondence problem* (the problem of what goes with what), which is just another sort of re-identification. Of course if the thing being tracked disappears and reappears there is the question of whether it is the same thing and if so in virtue of what properties it counts as being the same individual thing. The answer we will give is that in the cases we are interested in, it is the nonconceptual mechanisms of the encapsulated visual module that determine whether or not two tokenings are tokenings of the same individual thing, and it does so, I will claim, without benefit or the heavy equipment of concepts and without powers of reasoning. What properties of the individuals it uses is a separate empirical question to which we may or may not be able to provide a general answer, but the claim is that the properties that determine token sameness or numerical identity may act in a causal manner without themselves be represented. Consequently much will hinge on the mechanism of tracking and on the empirical properties of this mechanism, properties that we discover with experimental research. It is this sort of tracking, carried out by an encapsulated perceptual system (which I will refer to as “early vision”, after David Marr), that I want to discuss now. My examples will be from vision because more is known about vision (and my own experiments have involved the visual modality) but we will see later that the same story applies equally to (most) other modalities.

Before presenting the experimental demonstrations I might note that the role of tracking has been recognized by a number of people. The place of tracking in the metaphysics of objects was recognized by (Smith, 1996) and its role in demonstrative identification has been spelled out by philosophers like Gareth Evans (Evans, 1982). The importance that Evans places on tracking is illustrated by his insistence that “The fundamental basis .. of a demonstrative Idea of a perceptible thing is a capacity to attend selectively to a single thing *over a period of time*: that is, a capacity to *keep track* of a single thing over a period of time... (p 175)”. Two further points that this quotation does not address, and which I will argue are equally central is that (1) such a capacity to keep track must apply to more than one thing at a time and (2) the process of keeping

track does not rely on representations of any of the things' properties, including their locations. These two additional points, to which I have already alluded in the previous chapter, are illustrated by experiments in Multiple Object Tracking (MOT) to which I now turn.

## 2.2 Indexes and primitive tracking

In the last chapter I sketched a theory called Visual Index Theory (or FINST Theory) which assumes a finite (in fact, numerically small) capacity for indexing certain kinds of individuals. I suggested that the “indexing” process might be viewed as a process of demonstrative identification and reference, a suggestion that I take up again in this and the next chapter. Such a process could serve the important function of anchoring our perceptions to the world by allowing us to bind a small number of arguments in mental representations to individual distal objects. The existence of such a mechanism suggests that we ought to be able to keep track of a small number of moving items under conditions where we do not (or could not) encode properties that uniquely described the individual tokens. This observation led us to develop an experimental paradigm called *Multiple Object Tracking* (MOT) which has now been studied in hundreds of different experiments and has led to many surprising findings. These findings have far reaching implications for understanding individuation and other philosophical problems and so I will devote most of this chapter to describing the experiments and to discussing their implications.

Suppose we ask a person to keep track of several moving targets, such as small disks or squares, under conditions where no current property can uniquely identify these targets and distinguish them from identical-appearing moving nontargets. In these studies we refer to the items to be tracked as “target objects” and the nontargets as “distractor objects”, following the terminology used in the psychological study of attention. In a typical experiment, we arrange for the target objects to move unpredictably among a set of identical nontarget objects, and even to change their visible properties (such as shape or color) at random as they move. If we have a number of FINSTs available and if these become attached to the target objects, then it should be possible to keep track of which objects are the target objects – so long as there are not more targets than the maximum number of available FINSTs. Following this line of reasoning, we devised the Multiple Object Tracking (MOT) experiment as a test of this prediction. In this experiment (illustrated in Figure 2-1) a small number of target objects (usually around 4) are briefly distinguished from a number of visually-identical nontarget objects – typically by blinking the targets on and off a few times. Then all objects move around unpredictably on a screen, the targets traveling helter-skelter among the identical nontargets, for some period of time (say around 10 seconds). At the end of the trial subjects in this experiment must indicate in some way which objects had earlier been designated as the targets (they might do so, for example, by selecting them by clicking on them using a mouse, or by judging whether a single flashed object was a designated target).

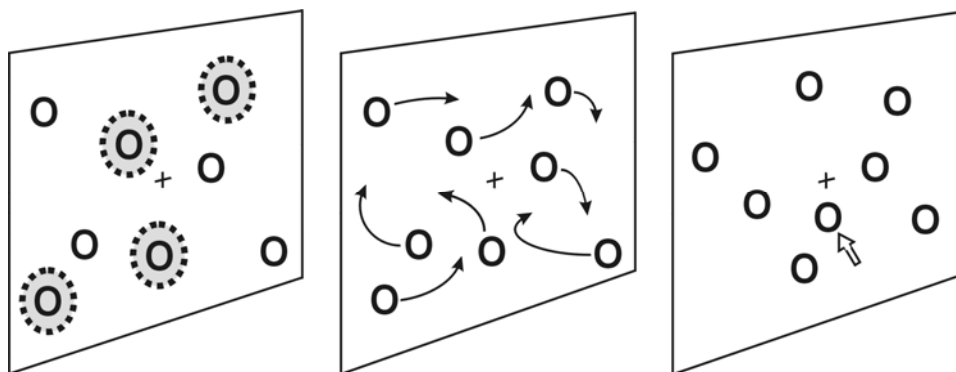


Figure 0-1: Sketch of the Multiple Object Tracking Experiment. Items shown with dotted circles around them are flashed a few times before they begin to move randomly. When they stop, the observer uses a computer mouse to indicate the items that had been designated as “targets” by having been flashed earlier.

While this may sound like a difficult task (especially when the nontargets and targets are allowed to move freely and to pass in front of each other in their travels), our volunteers find this task extremely easy to do for 4 targets and 4 nontargets moving at speeds that would be reasonable if the objects were real balls in some sort of field game (the actual speeds varied across time and for different experiments, but on average it was such that it would take a moving object from 4 to 6 seconds to cross the entire computer screen if it moved in a straight line – which it never does because of the random-walk algorithm used). The basic experiment has been replicated hundreds of times under many different conditions and in dozens of laboratories and performance of better than 85% in tracking 4 items is routinely observed (even young children aged 5-8 can track about 3 items Black & Pylyshyn, 2004; O’Hearn, Landau, & Hoffman, 2005; Trick, Audet, & Dales, 2003; Trick, Perl, & Sethi, 2005).<sup>10</sup> The experiment is most impressive when actually experienced. The reader is encouraged to examine recorded animated demonstrations of several versions of the experiment by visiting the web site: <http://rucss.rutgers.edu/faculty/pylyshyn/DemoPage.html>

Some of the conditions under which this task can be done will be discussed later since they illustrate a number of properties of the indexing and tracking process that are of theoretical interest; these include conditions in which objects change color and shape, disappear briefly behind an occluding surface or simply disappear from view entirely as though the observer had blinked. For now it will suffice to describe an analysis we carried out on the first published case of this experiment (Pylyshyn & Storm, 1988). In that study the trajectories of all moving objects were recorded so we were able to ask whether certain ways of keeping track of the targets could have been used by subjects in that experiment. For example, one possible way of tracking the 4 targets might be to encode and store their instantaneous locations and to continuously update them by moving attention to each in turn. This is not an unreasonable assumption since there is reason to believe that focal attention is required in order to encode the location of an object (Saarinen, 1996). Such a tracking strategy might proceed by recording in some form the

<sup>10</sup> Some published research includes (Allen, McGeorge, Pearson, & Milne, 2004; Alvarez, Arsenio, Horowitz, & Wolfe, 2005; Alvarez & Cavanagh, 2005; Alvarez & Scholl, 2005; Bahrami, 2003; Blaser, Pylyshyn, & Domini, 1999; Blaser, Pylyshyn, & Holcombe, 2000; Cavanagh, 1992; Cavanagh & Alvarez, 2005; Cavanagh, Labianca, & Thornton, 2001; Culham, Brandt, Cavanagh, Kanwisher, Dale et al., 1998; Fougny & Marois, 2006; Intriligator & Cavanagh, 2001; Jovicich, Peters, Koch, Braun, Chang et al., 2001; Keane & Pylyshyn, 2006; Ogawa & Yagi, 2002; O’Hearn, Landau, & Hoffman, 2005; Pylyshyn, 1989, 1994, 2004; Pylyshyn, Burkell, Fisher, Sears, Schmidt et al., 1994; Pylyshyn & Storm, 1988; Saiki, 2003; Scholl & Pylyshyn, 1999; Scholl, Pylyshyn, & Feldman, 2001; Scholl, Pylyshyn, & Franconeri, 1999; Sears & Pylyshyn, 2000; Slemmer & Johnson, 2002; Suganuma & Yokosawa, 2002; Trick, Perl, & Sethi, 2005; vanMarle & Scholl, 2003; Viswanathan & Mingolla, 1998, 2002; Yantis, 1992).



locations of all targets at the start of the trial (while they are visibly distinct) and then moving attention sequentially to each target in turn during the tracking process using their recorded locations. Because the objects are moving during this updating process the right object might not be at its stored location any longer. But we might nonetheless persevere and pick the nearest object to the recorded location and assume that this object is the target in question. Continuing with such a strategy, we might then update the object's stored location (assuming it was a target) and continue visiting and updating the locations on the list of presumed targets until the end of the trial. Using published estimates of the speed of attention movement we tested this location-updating hypothesis on the actual trajectories used in the experiment (note that this is a conservative test since it ignores the significant additional time that it takes to encode target locations, as well as to disengage and reengage focal attention, Danckert & Maruff, 1997). Simulating this strategy on the actual trajectories of objects used in our experiment yields a predicted performance of only about 30% under the most conservative conditions – i.e., using the highest estimates of attentional speed reported in the literature and even considering the possibility that not just location, but also speed and direction of each target is also encoded to enable some degree of prediction of the targets' location. This is far from the 87% performance we actually observed with our volunteer subjects (Pylyshyn & Storm, 1988). Thus we concluded that targets were not tracked by using their encoded and continuously updated locations.

Other more recent studies also showed that making every object a different color or size or shape does not help tracking (Dennis & Pylyshyn, 2002) nor does changing the objects' colors and shapes randomly during tracking make it worse. In fact subjects are unaware of changes in objects' properties during tracking (Bahrami, 2003; Scholl, Pylyshyn, & Franconeri, 1999). Thus we concluded that what makes it possible for objects to be tracked in MOT is the existence of a mechanism such as the one we long ago wished we could assume when we envisioned the Plastic Man fingers, the mechanism we call a FINST or Visual Index.

## 2.3 What goes on in MOT?

### 2.3.1 FINSTs and Object Files

A convenient way to tell the story of what goes on in MOT is in terms of Object Files (Kahneman, Treisman, & Gibbs, 1992). Object File Theory was developed independently of FINST theory and came out shortly after the initial FINST ideas and MOT were published. It bears a close affinity to our work – even though the Object File research relied on a very different experimental paradigm. But Object Files provide a useful way to think about how FINSTs get used in analyzing a scene and I will refer to them later in describing the relation between these approaches and the important philosophical work by philosophers who have written about individuals.

We can think of an Object File as a way for information to be associated with objects that are selected and indexed by the FINST mechanism.<sup>11</sup> When an object first appears in view (or attracts attention because it blinked) a file is established for that object. Each object file has a FINST reference to the particular individual to which the information refers. So, for example, in Figure 2-2 the file labeled *Object #27* contains some information about that particular object – e.g., that it is round and green and so on – whatever properties have been noticed about the object. It thus typically contains information about that object that was true in the past – such as that at some time in the past it was blinking and therefore was designated as a target in MOT.

---

<sup>11</sup> In contrast, (Kahneman, Treisman, & Gibbs, 1992) view a FINST as the initial state of an Object File, before any information is filed in it.

The FINST index mechanism is what allows this information to be associated with the *same token element* over time without requiring the object to be reidentified – the identity is maintained by the FINST mechanism that is built into the visual system. Notice that it need not (and most likely does not) have information about the properties that caused the index to be assigned or caused the object file to be created, nor does it necessarily contain information about which properties allow the individual object to be tracked. That takes place within an encapsulated part of the visual system and is a nonconceptual and causal processes (I will come back to this point in the next chapter). What the object file contains is just what was conceptually encoded, in particular the one-place predicates that pertain to that object. (The fact that it is conceptual does not imply that it is either conscious or involves concepts of the sort that a person could have – they could be subpersonal concepts that describe the state of proximal patterns).

For purposes of exposition, an illustration of what the FINST indexing mechanism does is shown Figure 2-2 below. It shows how FINSTs provide a reference link from an object file to an object in the distal environment and that it does so in response to a causal/informational event in the world that captures the FINST reference tokens. Clearly this leaves many details unspecified, many of which will be described later in this book.

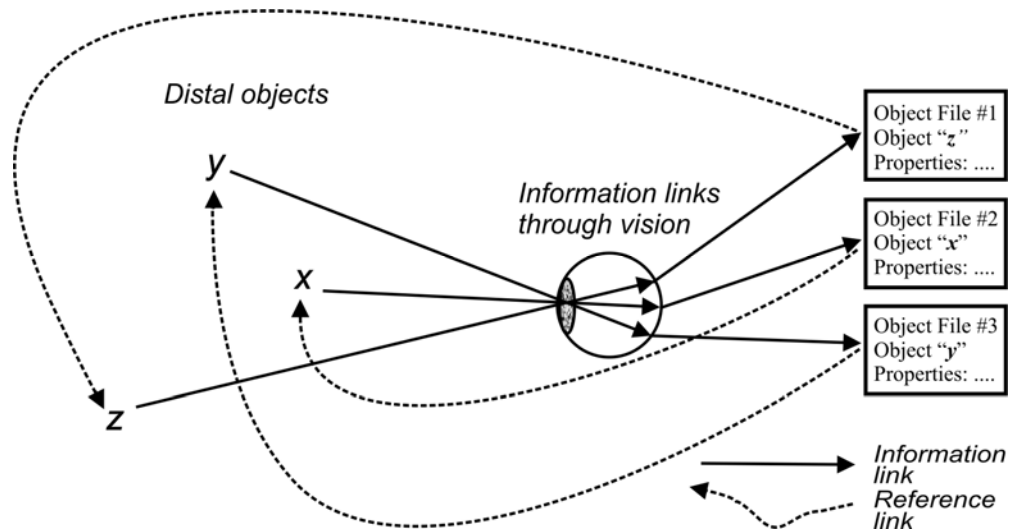


Figure 0-2. Sketch of the structure of the FINST mechanism and its relation to Object Files and to things in the environment. Notice that Object Files may be empty, as they appear to be in MOT.

### 2.3.2 The explanation of tracking

I have already provided, at least implicitly, an account of MOT in terms of FINST indexes and Object Files, but it will help our future discussion if I do it again in a slightly different way, this time emphasizing the more general implications of this way of looking at things. Here is the alternative account.

When a set of visual objects is blinked, each of them automatically captures an index, up to the maximum number available (around 4 or 5), presumably according to some priority scheme based on objects' salience. When an index is captured, an object file may be created for the object that captured it. The file is initially empty, and may remain empty, depending on whether there is an opportunity or a reason to enter object-specific information into it. As the objects move around, the object file remains attached to its respective object, so long as the object remains in view (actually a bit longer due to the inertia of the sensory system, but that is a

separate story I will not go into here). At the end of the trial when the object stops moving, various scenarios are possible. The simplest is that each indexed object is visited with focal attention and the subject moves the cursor to it (recall that “move attention to X” for an indexed object X is a primitive operation in our scheme, meaning that any explanation of how it works lies outside our present level of description, presumably falling under an architectural or neuroscience vocabulary).

Since it appears that nothing is stored in the object files under typical MOT conditions this suggests that targets are not being picked out under a description – they are not picked out as things that have certain properties or satisfy certain predicates. I have made much of this observation, interpreting it to mean that the FINST link is nonconceptual. But if it is nonconceptual, how can it re-identify an object as one that had earlier been selected as a target? How can it track an individual without conditions of individuation and without the concepts of object and of identity? And how can FINST indexes allow one to subitize, given that counting requires sortal concepts. This is where I believe Indexes are of interest to philosophy and I will return to these points later.

## 2.4 Other empirical and theoretical issues surrounding MOT

### 2.4.1 Do we track by keeping a record of locations?

We saw in section 2.2, keeping track of objects in MOT does not depend on recording and updating their locations. But at the same time, location is not irrelevant. After all it is because objects are at different locations that they are considered to be different objects in the first place. And it may be that even though a serial scanning and location-updating process cannot explain MOT performance some record of their location may be retained. It seems at least that when tracked targets disappear there is a record of where they were when they disappeared (Keane & Pylyshyn, 2006; Scholl, Pylyshyn, & Franconeri, 1999). But a study by Erik Blaser (Blaser, Pylyshyn, & Holcombe, 2000) shows that it is possible to track objects (or at least to track one object and ignore another) when the objects in question always occupy the same spatial location and move only in property space.

*Tracking through “property space”: The Blaser study.* Earlier I suggested that people track moving objects without encoding and appealing to any of their properties, including their locations. I will be returning to this claim in subsequent chapters for it represents a departure from one’s intuitions as well as from the received wisdom. But I want to provide one final example of a more general kind of “object” for which tracking clearly does not depend on keeping track of an object’s spatial location, because its location does not change during a trial. One might say that the object is tracked as it moves continuously through a property space.

The stimuli for this experiment consist of two rapidly-alternating superimposed patches of sign-wave gratings (these patches, called Gabors, fall off in intensity towards their edges according to a Gaussian profile, which gives them some desirable technical properties that are not relevant in the present context). Under these conditions, the superimposed Gabors looked like gratings painted on two transparent surfaces. The gratings varied randomly and continuously over time in their spacing (or spatial frequency), their color (which varied from red to black) and their orientation. Observers were asked to select a particular specified grating and to “follow” or track it over a period of time, indicating at the end of the trial which one they had been tracking. An example of the “trajectories” through feature space of these two objects are

shown in Figure 2-3 and the actually display sequence (without color) is also shown in Figure 2-3. Subjects were also given other less obvious tasks to perform to show whether they had successfully tracked a particular one. For example, they were asked to make a pair of judgments about two changes that occurred on these gratings (small sudden jumps in color or spacing). We found that people could not only track one of these spatially-fixed patches but they also showed evidence of treating them as single objects (e.g., two judgments were faster if they pertained to the same “object” than when they involved different objects, a technique for showing the object-specific nature of attention, to which I will return in Chapter 3). Thus it seems clear that people can track objects even if the objects are not moving through space, and therefore under conditions where their spatial locations do not serve to individuate them.

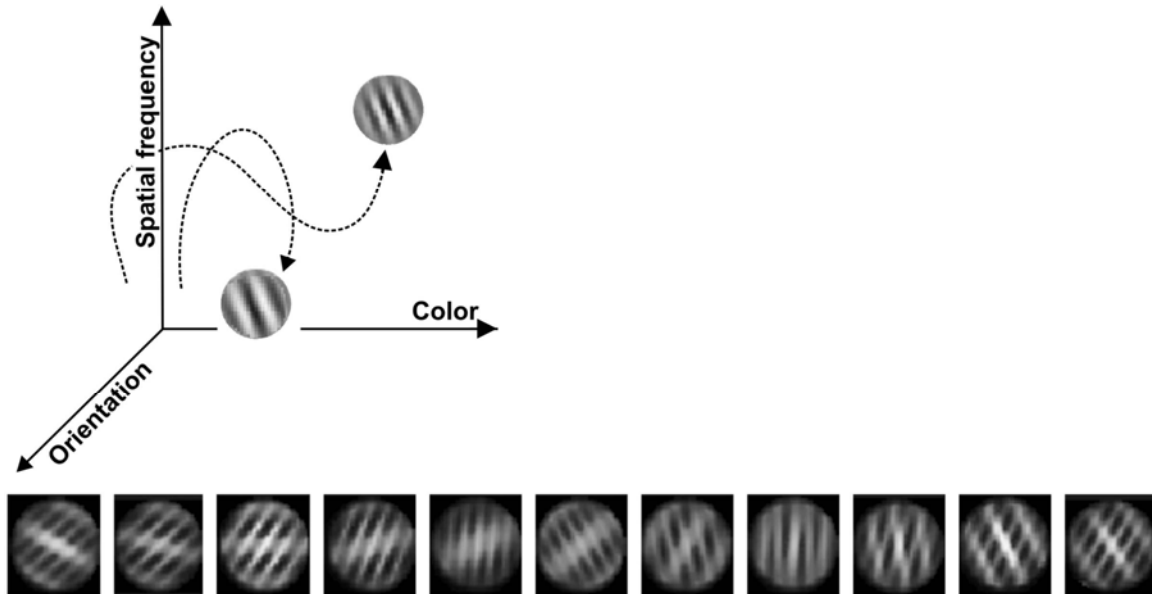


Figure 0-3. The bottom strip shows an example of the sequence of displays used by (Blaser, Pylyshyn, & Holcombe, 2000) to demonstrate tracking through “feature space” without motion in real space (the original displays were in color). The figure at the top shows a typical trajectory that each of the two “objects” follows through feature space.

## 2.4.2 Can we select objects voluntarily?

*Top-down and bottom up (or interrupt-driven vs test-based) selection.* I have assumed that selection of items is data-driven. In other words, properties of the objects capture or *grab* indexes on the basis of some item-specific properties that set priorities for indexing. But many investigators who have used the MOT task assume that participants in these experiments can choose when and whether to index particular items. In fact in some cases targets are indicated by color differences rather than by flashing. The reason I have assumed that indexes are basically captured by certain objects is theoretical. If indexes are the most primitive preconceptual contact that the mind has with things in the world, then the visual system is not in a position to decide what to index.

Consider what is involved in this issue. In (Pylyshyn, 1984, Chapter 6) I discuss the first line of causal contact between the world and mental representations – the mechanism that converts patterns of energy arriving at the sensors into symbolic codes. Such a mechanism is generally referred to as a *transducer*. To prevent a hypothesized transducer from subsuming all of perception (as happens in James Gibson’s “Direct Perception”, see Fodor & Pylyshyn, 1981),

and thereby losing its explanatory value, severe constraints must be imposed on how any hypothesized transducer can operate. Among these constraints is the requirement that a transducer (as opposed to the entire perceptual system) be *data-driven* – that its operation must depend only on incoming information as opposed to depending on the visual/cognitive system. A data-driven mechanism is triggered by (or is responsive to) the arrival of an information-bearing signal from the world rather than being responsive to a query posed by the visual system. In computational terms it is based on an *interrupt* rather than on a *test* applied by the perceptual system (e.g., a conditional “if-then” operation). We sometimes refer to this as bottom-up as opposed to top-down control. But it should be noted that regardless of the details of the empirical data on the operation of a system that represents information about the world, such a system *cannot be entirely top-down*. At some point the world must cause certain effects on the perceptual system – which is to say, it must impose itself onto the system willy-nilly. Even a top-down system that checks for the presence of certain potential or hypothesized properties can only check for information that has made it into symbolic form – i.e., information that has been transduced in a data-driven manner. A system cannot literally reach out into the world to check on properties unless the properties are among those that are able to affect the system causally, which means in a bottom-up manner. This is also true in the case of real computer systems as well; any if-then test that is computed can only be do so on some already-transduced signal (often a flag that is set by an interrupt).

*Voluntarily enabled selection.* Not surprisingly, we were able to demonstrate experimentally that observers can voluntarily select targets to track, even when the selection criteria are ones that themselves require focal attention (Pylyshyn & Annan, 2006). For example, when the items are numbered, people can select items numbered 1-4 and track them, while ignoring items numbered 5-8. Then we looked more carefully at how long it took to make the selection of targets from among the 8 disks. In one of our conditions a bar flashes on the disks that subjects had to track. We compared how long this flashing cue had to be present compared to how long it had to be present if the task was to track the disks that did not have a flashing bar (the complement set). Assuming that the flashing bar automatically attracts FINST indexes we expected that for a given flash duration the track-flashed condition would be better than the complement (track-unflashed) condition, which is exactly what we found. We also presented a condition in which a horizontal bar flashed on 4 of the 8 disks and a vertical bar flashed on the remaining 4 disks, and vice-versa. The task was to track the disks that had horizontal bars (or sometimes the ones that had vertical bars) and ignore the disks that had vertical bars (or alternatively the ones with horizontal bars). Again we found that when the selection criterion required attention (track flashed-vertical among flashed-horizontal) the cue had to be present for longer than when the selection criterion was automatic (track disks with flashed bars among disks without flashed bars). Moreover, when the number of targets was increased (from 3 to 4 and 5) the amount of additional time required was also greater. All these results suggest that while we can select items voluntarily, the selection takes more time because the items have to be visited one at a time with focal attention.

This is an instance of a general principle. Automatic functions often require, or at least can be affected by, voluntarily applied enabling conditions. For example, perceptual phenomena typically depend on direction of gaze – few phenomena work when the critical part of the stimulus is not foveated. Often the phenomena depend not on gaze but on focal visual attention. For example, the automatic perception of a line drawing as a 3D figure (as in the Necker cube) is affected by the locus of attention (and may account for why some people mistakenly feel they can influence which version of the ambiguous cube they see merely by a act of will, e.g., Churchland, 1988). Some of these affects of focal attention are surprising. For example, certain

effects of masking of one pattern by another presented briefly at the same location (and for slightly longer than the original pattern) are only observed when the pattern is attended (Di Lollo, Enns, & Rensink, 2000). The automatic stereovision system can also be sensitive to the locus of observers' attention (Frisby & Clatworthy, 1975). It is also known that certain features can be filtered by early (preconceptual) stages of perception, as though a "switch" can be set to prefilter certain properties (the earliest theories of attentional selection were in fact explicitly filter theories, e.g., Broadbent, 1958). Certain automatic effects in auditory phonetic recognition (such as the categorical perception of phones) occur only when observers are set for a linguistic stimulus, as opposed to music or an arbitrary noise (Best, Studdard-Kennedy, Manuel, & Rubin-Spitz, 1989). These are not cases of the cognitive penetration of perception (of the sort discussed in Pylyshyn, 1999) because it is not a rational content-dependent influence. Rather it is a case where the functioning of an automatic mechanism (or perceptual module) can be enabled by a voluntarily controlled setting. What happens with voluntary selection of targets in MOT is, I believe, a case of an automatic data-driven function being enabled by a voluntary act of focusing attention. In vision, automatic processes such as "singleton" feature selection may depend on the local uniqueness of certain properties. What focal attention may do is define an area over which the uniqueness is computed. This proposal is consistent with evidence that when attention is directed to a particular location, a unique target ("singleton") near the site of the focal attention is more likely to capture attention than one more distant (Mounts, 2000). In any case this suggests a plausible account of how voluntary selection can occur in a way that is compatible with the independently motivated requirement for an automatic mechanism that captures reference indexes ("grabs FINSTs").

The empirical evidence concerning individuation and tracking is rich and often surprising. Here are a few findings that continue to challenge our understanding of what goes on in tracking and which also have implications for the philosophical issues concerning nonconceptual representation, individuation and the tracking of identities.

### 2.4.3 Tracking without keeping track of labels

One early finding concerns the failure to recall the *identity* of targets while tracking. Our story of how we track multiple moving objects in MOT using FINST indexes claims that an index keeps referring to the same individual object throughout the experimental trial. In doing so it keeps track of the identity of each individual target during this period. In other words it knows that a particular individual object  $X_i$  is a target because it was visibly distinguished as a target at time  $t_1$  and is the same individual now as it was at time  $t_1$ ; therefore it must be a target. If this story is correct then if the object had been labeled with label  $L_i$  at the start of the trial and if we can show that, under conditions of the experiment, subjects can recall the four pairings  $X_1 \leftrightarrow L_1$ ,  $X_2 \leftrightarrow L_2$ ,  $X_3 \leftrightarrow L_3$ ,  $X_4 \leftrightarrow L_4$ , then subjects should be able not only to identify the targets as targets, but they should be able to provide their labels. But in fact people are very bad at recalling targets' labels even when they correctly identify them as targets (labels were either numbers affixed to the disks or just the distinguished locations of the disks in the corners of the display at the start of each trial, for details see, Pylyshyn, 2004). How can that be? In order to track on object it is necessary that its identity be traced back to a prior state when it was known to be a target, which ultimately means it must be traced back to the beginning of the trial when it was visibly distinguished as a target. So how can one not also remember its name which represents a very small additional memory load that we can show is well within the subjects' memory capacity. The latter is shown by the fact that the disk-name pairs can be recalled nearly perfectly

so long as the tracking and label-recall involves different objects.<sup>12</sup> (An example of the sort of display used can be viewed at: <http://ruccs.rutgers.edu/finstlab/demos.htm>).

According to the above logic, if all targets were tracked perfectly we would expect the labels to be recalled perfectly as well. But tracking is not perfect, and switching the identity of a target with that of a nontarget is recorded as an error in tracking, while switching the identity of a target with that of another target does not affect the tracking score. Thus the reason for the failure to keep track of labels might be traced to the fact that distribution of errors is inhomogeneous. Such an account is proposed in (Pylyshyn, 2004). If the likelihood of a target being mistaken for another target is higher than the likelihood of a target being mistaken for a nontarget this would explain the observed difference between tracking performance and labeling performance. But why should the likelihood of these two types of confusions be different – why should targets be more often confused with other targets than with nontargets? The answer I suggested in that paper is that during tracking an attempt is made to eliminate the items which might cause confusion (i.e., the nontargets) by a process of inhibition. Such a process has been observed in many situations and is thought to be invoked under conditions where there is a task-specific need to keep certain items in a display from attracting attention (or, in our terms, from grabbing an index). While such an assumption might explain the asymmetry, it raises an even more interesting issue: how can items that are not being tracked be inhibited unless either everything that is not a target is inhibited or unless the nontargets are also being tracked? This question is discussed below because it has far-reaching consequences for our view of individuation and tracking.

#### 2.4.4 Nonconceptual individuation without reference?

As many philosophers have pointed out, true individuation requires concepts and conditions of identity. But the earlier discussion suggested that some functions that require concepts in general may have nonconceptual solutions, at least in particular cases. One such example concerns computing identity in apparent motion and stereovision when solving what is referred to as the “correspondence problem.” The correspondence problem is the problem of computing the identity of distal causes from distinct proximal (i.e., retinal) perceptual clusters. In other words it is the task of treating two temporally- or spatially-distinct tokens of local features as arising from the same remote object. This is a ubiquitous problem in vision science and is especially clear in computing apparent motion (where temporally and/or spatially disparate objects or flashes are seen as a single moving distal object) and in stereovision (where two distinct objects, one on each retina, are seen as arising from the same feature in a 3D scene, a prerequisite for computing their retinal disparity and thus their distance in depth). These (and other) cases of correspondence computations exhibit important properties which suggest that they are computed by special mechanisms in the encapsulated early vision system that do not use conceptual properties of the object tokens. Moreover there is reason to think that these mechanisms are different from the FINST mechanism in a way that bears on the puzzle of inhibition of moving nontargets raised above.

---

<sup>12</sup> For those interested in the experiment, it goes something like this (details in Pylyshyn, 2004). A (static) set of randomly placed numbered disks is presented and subjects told to try to remember their numbers. This is followed by the usual tracking task involving *different* disks without labels. Then the first set of disks (the ones that had not been tracked) is presented again in their same fixed locations – this time without labels – and subjects are asked to recall their labels. In this condition recall of the labels was nearly perfect, showing that the tracking task in and of itself does not interfere with the simple recall of 4 pairs of objects and their labels.

Neither tracking nor the correspondence computation is very sensitive to the local featural properties (e.g., color or shape have little effect on the correspondence matches in apparent motion, Kolers & von Grunau, 1976) of the objects in question (with perhaps the exception of a few properties such as their spatial frequency or the polarity of the luminance of the objects – black on gray versus white on gray). In both cases the process appears to work best when there are distinct individuals, as opposed to smoothly varying luminance functions. But a major difference between these processes is that the correspondence computation in apparent motion and stereovision does not appear to be limited in capacity the way MOT is. In fact correspondence may be established more readily when there are more items to be placed in correspondence. For example displays that contain more dots are more easily interpreted in the so-called kinetic depth effect (Wallach & O'Connell, 1953), an apparent motion display in which perception of a three dimensional surface in motion appears from a properly arranged sequence of dots, or in random dot stereograms (Julesz, 1971), in which a field of random dots, presented to the two eyes so that there is a relative location disparity between them, produces a perception of a dotted surface in depth. In these cases it appears that computing correspondence between pairs of dots is not a numerically limited process the way that visual indexing is. This suggests that individuating objects and putting pairs of them in correspondence may be distinct from individuating and referring to objects, a process that must occur for successful tracking in MOT. Recall that at the end of an MOT trial, subjects must select particular individual objects as targets, for example by moving attention to them and then moving a cursor to each in turn in order to indicate that they were the targets. In the correspondence examples no information about which object was which is needed over and above what is required for pairing them. Moreover the pairing can be done on the basis of spatially local information – it is a computation carried out with what is called “local support”.<sup>13</sup> This suggests that we may need to distinguish the process of feature-clustering and individuating from the process of picking out and referring to objects.

The need for such a distinction has also come up in connection with some surprising findings in recent studies of MOT. Because these findings suggest that individuating and computing clusters of features that belong together is different from the process of tracking, I will describe one of these experiments for concreteness.

As mentioned in the previous section, we had reason to believe that target-target confusions (where the identity of one target is confused with the identity of a different target) is more common than target-nontarget confusions and that this might be due to nontargets being kept out of contention in MOT by a process known as inhibition. Inhibition is a quite general phenomenon in perception. For example it has been shown that objects that have just been attended are less likely to be attended again immediately after (i.e., within about 300 to 900 ms), even if attention is summoned there by a flash. This phenomenon, referred to as “inhibition of return” (Klein, 1988) is thought to help us in searching through a set of candidate visual objects (as in the “*Where’s Waldo?*” game). This kind of inhibition has also been shown to be associated with objects as opposed to (or in some cases in addition to) locations and to move with moving objects (Tipper, Driver, & Weaver, 1991). There is also some evidence that multiple moving objects might be inhibited in this way in the course of searching through them for a specified target (Ogawa, Takeda, & Yagi, 2002).

To examine whether nontargets were inhibited in MOT we carried out a number of experiments (described in Pylyshyn, 2006) using a method that involves presenting a small probe

---

<sup>13</sup> The criterion of local support is important insofar as functions meeting this criterion may be computed by a network of cellular automata which, in turn, makes them good candidates for being a function of early vision (Ullman, 1976).



dot during an MOT trial. Performance in detecting such a dot has been used as a measure of both attention and inhibition (Watson & Humphreys, 1997; Watson & Humphreys, 2000); performance in detecting the dot is better on attended items and worse on items that are being actively ignored. We presented probe dots not only on targets and nontargets, but also in various places in the empty space between them. We found that, as expected, detection of the probe dots is poorer on nontargets than on targets. But surprisingly, detection in empty space is about the same as on targets (when compared with a control condition in which subjects merely watched for probes without tracking). This result (which has recently been replicated Flombaum, Scholl, & Pylyshyn 2006) suggests a puzzling conclusion; namely, that the inhibition attaches to moving nontargets without affecting the space through which they move. In other words, it looks like inhibition is *tracking* the nontargets. But according to the present view, nontargets could not be tracked in MOT using FINST indexes because there are a limited number of these indexes and, by hypothesis, they are all being used in tracking the targets. So how could the inhibition “track” along with the nontargets?

While the jury is out on whether there is an alternative way to explain these findings (for example it may be that what is being inhibited is anything that moves but is not a target) let us accept the above analysis for the sake of argument since it raises an interesting possibility. It raises the possibility that keeping inhibition attached to moving nontargets involves the mechanism that is responsible for solving the correspondence problem – i.e., that it is carried out by a mechanism that automatically and nonconceptually puts together the sequence of proximal tokens into the percept of an enduring individual, as it does in apparent motion and stereovision. Unlike the FINST index, however, this mechanism merely collapses a sequence of time slices of proximal clusters but does not provide a reference to them. According to this analysis there are two distinct functions involved in tracking in MOT. One consists in individuating moving items and collapsing them over time and space according to the correspondence principles, and constructs perceptual enduring individuals. As we saw in the case of apparent motion and stereovision, this aspect is not numerically limited and is computed based on spatially local information (it follows the principle of local support). The second function is the one I referred to as demonstrative reference – it consists of providing a conceptually-unmediated way to refer to the individuals, using FINSTs. This function is numerically limited, allowing reference only to about 4 objects. The first function applies to all moving objects in the field of view; the second only applies to objects that have been selected as described in section 2.4.1 (i.e., the targets).

While this story remains provisional and somewhat speculative, it does suggest that an account of individuating and tracking may be more complex than most people have assumed, and that the question of the existence of nonconceptual counterparts to individuating and tracking may be both more nuanced and ramified – and also more interesting – than one might have expected. That’s why empirical evidence remains essential in developing ideas about how the mind connects with individual things in the world.

## 2.5 Infants capacity for individuating and tracking objects

Before discussing the philosophical implications of such experiments I should mention that the FINST framework, together with the idea of object files (Kahneman, Treisman, & Gibbs, 1992), has provided a way of understanding another sort of tracking – infants’ tracking of individual objects and their sensitivity to the cardinality of sets of objects that are moved about and hidden behind a screen. The ideas discussed above have been used in analyzing these fascinating findings (Carey, forthcoming; Carey & Xu, 2001). These studies used the “violation

of expectation” method in which infants’ looking times are taken as a measure of whether the infant’s expectation was violated by the display. The basic finding is that (contrary to Piaget’s claim that infants do not have object constancy at a young age) infants as young as 6 months understand that if you hide a toy behind a screen and then you place another toy behind the screen as well, then when the screen is removed there should be two toys (there is even evidence of object constancy in 3-month olds, Baillargeon & DeVos, 1991). This is shown by the finding that if there is only one toy when the screen is removed, the infant looks longer, which is interpreted to mean that the result was unexpected. The same is true when a toy is seen to be removed from behind a screen where two toys had been placed earlier – infants look longer if there are two toys rather than one. A large number of experiments have been carried out using not only the “looking time” method but other methods as well (e.g., after how many found items does an infant stop reaching into a bag looking for more, Van de Walle, Carey, & Prevor, 2000). They indicate that infants can respond to the cardinality of sets of 3 or fewer (see the review by Krojgaard, 2004).

Later studies showed even more subtle effects. For example, experiments showed that infants only respond correctly to the cardinality of two identical items if they saw both at the same time before they were hidden, but would respond to two different items if they were seen one at a time before being hidden (two identical items seen one at a time are interpreted as one item seen twice?). The findings keep refining the abilities of infants. For example, infants under 10 months of age were shown a red disk being removed from behind a screen so the infant could see it before it was replaced behind the screen. This was followed by showing a green disk being removed and then replaced. The finding was that in this condition infants expect two objects (two different colors = two individual objects). But interestingly, the infants were indifferent as to whether the two objects that they saw when the screen was removed were both red or one was red and the other green. In other words they used objects’ colors to individuate them and to infer that there were two, but they failed to use the objects’ colors in forming expectations and recognizing anomalies until they were over 10 months of age (Tremoulet, Leslie, & Hall, 2000) or in some cases 12 months of age (Xu, Carey, & Quint, 2004). Alan Leslie distinguishes between what he calls “individuating” and “identifying”. The latter, but not the former, presumably requires that the relevant properties be entered in the object files. This is very similar to the distinction I have been making between selecting and encoding, or between causal properties and properties that are represented.

More recently there has been evidence that infants *can* use other properties such as basic-level kinds (Xu, Carey, & Quint, 2004) and distinct faces (Bonatti, Frot, Zangl, & Mehler, 2002) in determining cardinality and in recognizing anomalies in the properties of items. What these studies confirm is that, as I claimed earlier, one must distinguish between properties that cause indexes to be grabbed, and properties that are represented (and stored in object files). The way properties get used in various tasks is rather complex as well. Studies have shown that infants do not treat piles of sand poured onto the stage as individuals: They do not respond to the cardinality of objects that were poured (Huntley-Fenner, Carey, & Salimando, 2002) – and, by the way, adults too are poor at tracking objects that move from place to place in an MOT experiment if the movement is like pouring or worm-like slinking (vanMarle & Scholl, 2003). Nor do infants respond to the cardinality of objects if they see the objects (made of blocks) taken apart and put together again (Chiang & Wynn, 2000).

Although the ideas of FINSTs and object files are used in explaining these results, some people have interpreted the demonstrations as showing that infants have the sortal concept *object* that they used to individuate (Xu, 1997) (but see the contrary opinion of commentators Ayers, 1997; Hirsch, 1997; Wiggins, 1997). The question of whether infants have the concept of object

is an interesting one, but it does eventually run into the need to ground that concept in experience (by “ground” I mean connect the concept with its instances, not *learn* or acquire the concept which may well be innate). For example, it has been suggested that the first sortal concept that an infant has is the concept *object* (as part of what is called their “core knowledge”) which is the concept of something that is “bounded, coherent, three-dimensional, and moves as a whole” (these criteria were introduced by Liz Spelke in, Spelke, 1990, so it is sometimes referred to as a “Spelke Object”). But of course if that’s what an object is for an infant, then infants must also have the concepts “bounded”, “coherent”, “three-dimensional” and “moves as a whole”, in which case Spelke Object could not be their first concept. If, on the other hand, Spelke Object is not defined in terms of other concepts then it cannot apply to all and only things that are bounded, coherent, etc. So what determines the extension of the concept? Once again, the distinction between causal properties and properties that are represented is crucial. If one takes the view proposed in this book, then the claim would be that infants start off with an architecture that determines their sensitivity to certain properties – including many abstract properties (such as “proto-cause” and perhaps something like “Proto-Spelke-object”) that are largely coextensive with the conceptual equivalents – but which cannot be identical to these conceptual properties. They are neither conceptual *definienda* nor *definienda*. They are constituents of thoughts only in the limited way that indexicals or names are constituents; they only contribute reference to the content of complex thoughts. By contrast, some properties do get encoded in the form of predicates, since predicates are properties of indexed things, so FINSTs are logically antecedent to predicates. In other words, Red is a predicate that can be bound to things selected by FINST indexes, which results in Red(F<sub>1</sub>) being stored in F<sub>1</sub>'s object file, i.e., the file corresponding to the demonstrative referent of index F<sub>1</sub>. I will argue that it is a general property of conceptualizations of the perceptual world that only indexed objects can serve as arguments of predicates and consequently only properties of selected objects are conceptually encoded.

## 2.6 Summary and Implications for the Foundations of Cognitive Science

### 2.6.1 Review: Nonconceptual functions and Natural Constraints

By hypothesis, indexing and tracking in the context of MOT are nonconceptual functions carried out within the modular visual system – a system that may or may not have representations (such as the primal sketch or the 2 ½D representation described by David Marr, Marr, 1982).<sup>14</sup> Yet indexing and tracking appear to contradict the prerequisites on individuals, particulars and identity discussed by Quine, Strawson and other philosophers. For Quine (Quine, 1960) you can’t think about particulars in the world without what he called “an apparatus of individuation”, by which he means sortal concepts and identity and maybe divided reference and tenses. Strawson also considers individuating and reidentifying to be essentially conceptual functions. He writes in his book *Individuals* (Strawson, 1959), “...the introduction of particulars ...involves a conceptual complication: it involves the adoption of criteria of distinctness and...

<sup>144</sup> For now I leave open the question of whether the early vision system that implements Indexes has representations of any kind. The evidence I have reviewed suggests that no representations of object properties are used in tracking. Moreover any representations in this part of the system are encapsulated from the cognitive system (Pylyshyn, 1999). The question of whether at this level it uses concepts is to some extent terminological: If it has concepts they are what some (Dennett, 1978; Haugeland, 1978) have called sub-personal concepts, not the sort of concepts that form part of the thoughts that we, as human agents, may have; not the sorts of concepts that we would recognize as *our* concepts if they were conscious. Such sub-personal concepts may, for example, be codes for proximal properties involved in perception, such as edges, gradients, or the sorts of labels that appear in early computational vision (Marr, 1982), or parsing trees in language understanding/generation.

criteria of reidentification for particulars of the kind in question. (p. 203)” (see also, Keane, 2004). And so it does. Early vision does not have the power of predication and does not have count nouns or sortal concepts or the capacity for past tense reference (also part of Strawson’s criteria for individuation) and consequently cannot identify or re-identify individuals as the same individual encountered previously. Yet under certain conditions (namely the conditions that allow indexing and tracking) FINSTs do allow us to individuate and even to “re-identify” certain individuals: They allow us to maintain the identity of tracked objects as enduring individuals and they even play a role in recognizing the cardinality of the set of objects when it is no greater than about 4 (though assigning an actual *number* to the set clearly requires the concept of number). Once we have indexed an object we can keep track of it (within some broad range of conditions). And with the mechanism of Object Files we can also accumulate past-tense information about the objects. And the visual system can do all that without the heavy apparatus that is required in principle.

How is that possible? It is possible because wired into our early vision system by evolution is a mechanism that provides a type of identity that is not conceptual. This should not be surprising since vision provides many nonconceptual functions that are, as a matter of contingent fact, extensionally equivalent to conceptual categories *in our sort of world*. For example early vision has mechanisms for constructing three-dimensional shape representations from 2D (retinal) images, despite the fact that the 3D to 2D mapping is in principle not reversible. This irreversibility had been taken by vision scientists over the past 30 years as proof that constructing a 3D representation could only be done framing hypotheses based on general knowledge of the world. But vision does this without concepts and without any knowledge of what makes something a real physical 3D object or what objects are likely to be in its visual field on any particular occasion (see the discussion in, Pylyshyn, 1999). It does so in virtue of constraints built into the visual system (other related constraints are discussed in Section 5.2). It constructs a representation that is generally correct given that the perceived world meets certain general constraints that tend to hold in our sort of world (or our ecological niche). By “our sort of world” I mean a world where, for example, our visual field is such that the vast majority of image features (features in the proximal or retinal image) have the property that close neighbors tend to arise from object-features that are at similar distances from the observer. The reason this constraint holds is that in our sort of world projections of continuous surfaces onto the retina tend to greatly outweigh in area the discontinuous edges. This depth-continuity constraint is used in determining which image features in one eye should be associated with which features in the other eye in computing stereo disparity. Similarly in our sort of world the majority of image features tend to arise from the surface of rigid objects and therefore tend to move together when the object moves. Consequently in solving the correspondence problem mentioned earlier, preference tends to be given to pairings that preserve such continuity, i.e., where the distance and direction of the correspondence pairs are similar to those of its neighbors (a natural constraint used in the model of apparent motion developed by Dawson & Pylyshyn, 1988); see also (Dawson, 1991). Another powerful constraint is the fact that when edges are aligned or form a vertex in an image, then barring accidental coincidences such edges are also aligned or form a vertex in the distal scene. These “nonaccidental” image properties are used in the recognition of scene properties because the likelihood that they arise by chance is very low in “our kind of world” and when they do arise they can easily be diagnosed by a small movement on the part of the observer. Also in our sort of world the light tends to come from above so shadows fall below the features that cast them, a fact that influences whether certain shadings are interpreted as convex or concave (as mounds or holes) (for these and other examples, see Pylyshyn, 2003, chapter 3). Other such constraints are discussed by (Hoffman, 1998) who refers to them as “intelligent” solutions to the problem of vision, though they are all applications of general

constraints that are built into the visual system as opposed to inferences from conceptual knowledge. They are all examples of how a nonconceptually-based system, operating in a well-defined niche (which may include most humanly-habitable worlds), can mimic a system that uses concepts.

Although geometrical properties are the more natural ones if we are considering natural constraints, vision also provides a nonconceptual mechanisms for much more abstract properties. For example it provides a mechanism, available even in 6 month old infants, for recognizing causal sequences (this is the mechanism that was initially studied in adults by Michotte and explored in infants by Alan Leslie, e.g., Leslie, 1982). Luckily for us, our visual system (and presumably the same is true for other perceptual modalities) provide us with functions that are not just approximation to these essentially conceptual skills, but ones that that almost always work in our kind of world. And the ones that are internalized are not just the more frequently seen patterns. Perhaps surprisingly, commonly observed physical properties, such as simple prohibition of interpenetration of solid objects, don't seem to be internalized as natural constraints. Alan Leslie discusses an example (Leslie, 1988), based on the Pulfrich Double-Pendulum illusion (see Wilson & Robinson, 1986), that makes this point in a dramatic way. Suppose two side-by-side pendulums are set swinging out of phase (so that one is just starting its motion to the left when the other is starting its swing to the right). If they are viewed through neutral-density (gray) filters such that the filter over one eye is darker than the one over the other eye, the pendulums will be seen to swing in a circular (or elliptical) path in depth rather than swinging back and forth in a plane. The reason for this illusion is thought to be that the weaker signals going to the eye with the dark filter result in slower visual processing in that eye which, in turn, results in the apparent circular or elliptical path. In Leslie's example, he notes that the consequence of this Pulfrich illusion is that the pendulums are seen as going *through* one another. Unlike many illusions this one can be seen in full close-up reality and there is no doubt to the observer that the bottles at the end of each wooden pendulum are solid sand-filled detergent bottles, yet they are clearly seen to interpenetrate one another. Another example is the Ames trapezoidal window demonstration. When a trapezoidal frame is rotated about a vertical axis it is seen, not as rotating, but as oscillating back and forth. This illusion is quite powerful. But if a solid bar is tied to the rotating axis, at right angles to the axis, it will be seen (correctly) as rotating in one continuous circle. But now the two percepts together result in seeing the bar passing through the trapezoidal window even though both are solid. So it seems that although many natural constraints are built into the visual system, they tend (with a few exceptions like causality) mostly to be optical-geometrical rather than physical-mechanical constraints. It could have been otherwise (and if it had, it would still receive a perfectly logical evolutionary explanation), but as a matter of fact, it isn't.

This idea of natural constraints is quite powerful. Yet I will suggest in Chapter 4 that it does not seem to have much of a role in explaining how mental imagery works. As you watch your mental image rotate, slide, twist, expand, get cut up, superimposed on other images, and so on, what you see is just what you expect to see – exactly what you intended to happen. Nobody has produced examples of phenomena that seem to reveal constraints of the cognitive architecture that are specific to mental images, the way such constraints are revealed in vision.<sup>15</sup> Part of the reason, I believe, is that mental images are really not images so much as types of thoughts and thoughts do not easily reveal built in constraints. This is not to say that there aren't any – there

---

<sup>15</sup> "What?" I hear you say. How about imagining an object described by basonic string theory that has 26 dimensions? "Okay," I reply, "as soon as you show me what it looks like I will imagine it, because that's what imagining, in the sense of "visualizing" means, it means having a visual experience." There are of course capacity limitations in mental imagery, but these appear to be the same sorts of limitations that apply elsewhere in cognition (e.g., limits on the number of chunks or units that can be attended at any one time).

must be thoughts that a dog or a cat or a chimpanzee (or, for that matter, a cosmologist) can have that you can't, but it's not easy to imagine one (exactly – because that would require imagining what for you is unimaginable!). As (Fodor, 1980b) has colorfully put it: “*From in here* it looks as though we're fit to think whatever thoughts there are to think... It *would*, of course, precisely because we *are* in here. But there is surely good reason to suppose that this is hubris bred of an epistemological illusion. No doubt spiders think that webs exhaust the options (p333).”

### 2.6.2 Summary: Why are FINSTs needed?

I want to end by providing a quick preview of why the FINST mechanism I have described is important to cognitive science. FINSTs give us nonconceptual access to what I have called a *thing* or a *sensory individual* or *visual object* or, in the context of FINST theory I sometimes refer to the things that FINSTs refer to as FINGs. Because the representation is not conceptual, these sensory individuals are not represented *as* objects or as X's for *any* possible category X. They are just picked out transparently by a causal or informational process without being conceptualized as something or other. Early vision picks out and indexes a small number (4 or 5) of such sensory objects, roughly the way you might pick out a fish by placing a baited hook in the water – it happens primarily at the initiative of the objects; we say it is data-driven. Selecting a subset of individuals in this way allows the cognitive system to encode and accumulate predicates about them, which, in the case of unary predicates it places in the *Object File* for that visual object. And as long as there are not more than 4 or 5 of these individuals the visual system can treat them as though it had a concept of “individual object”. But it's important to remember that the early vision system does not have the concept “object” or any other concept (except maybe the demonstrative token *this* if you want to count that as a concept on the grounds that it can partake in thought) so it cannot infer from the fact that it is tracking X that X cannot go through walls or even that there is an X, the way you could if you were tracking something as a planet or a baseball (from which you could infer that there is such a thing as a planet or a baseball). Part of our project in postulating FINSTs is that we need to get our cognitive mind to select something before we cognize the something in terms of a concept, such as the object X. Otherwise the question arises: In virtue of what is that an X? – which starts you on the slippery slope of asking the same question about each of the conditions for X-hood. Q: In virtue of what is that an object? A: Because it has mass and moves through smooth continuous trajectories in space. Q: And what does it mean that it has mass...? That requires that have the concept *mass*, and how do you ask about its trajectory unless you have the concept *trajectory*, and so on recursively. The recursion has to end somewhere and where it ends might as well be something for which you have at least some independent motivation and, if you are lucky, some empirical support. That's where FINSTs come in.

Another way to say why a mechanism such as a FINST is important is that even if the early-vision system somehow *had* the full conceptual apparatus of individuation and identity it still could not use that apparatus to connect to the world unless at some stage along the processing chain the leap from things in the world to concepts happens causally. You can't go on building concepts from other concepts without eventually bottoming out in a purely causal/informational connection (including, possibly, connections with the cognitive architecture). All of this may seem like heresy in the face of the important philosophical work of Strawson, Quine and others, and its ramifications could be far reaching. While philosophers may continue to understand “individual” as a construct that essentially involves concepts, it should not be forgotten that it must ultimately rest on a nonconceptual mechanism such as the one I have been describing. In discussing the potential value of FINST indexes to philosophical issues a few years ago I made the following remark to which I still subscribe:

“While it is clear that you cannot individuate objects in the full-blooded sense without a conceptual apparatus, it is also clear that you cannot individuate them with only a conceptual apparatus. Sooner or later concepts must be grounded in a primitive causal connection between thoughts and things. The project of grounding concepts ... in perception remains an essential requirement if we are to avoid an infinite regress. Visual indexes provide a putative grounding for basic objects – the individuals to which perceptual predicates apply, and hence about which cognitive judgments and plans of action are made... Without such a preconceptual grounding, our percepts and our thoughts would be disconnected from causal links to the real-world objects of those thoughts. With indexes we can think about things ... without having any concepts of them: one might say that we can have demonstrative thoughts.” (Pylyshyn, 2001, p154).

## Chapter 3. Selection: The key to linking representations and things

### 3.1 Selection: The role of focal attention

We have been discussing the connection between the world we perceive and mental representations. This topic has a way of returning again and again to the notion of selection. *Selection* is a central topic in contemporary cognitive science and, as we shall see, it is also the place where empirical cognitive science comes in contact with a series of problems in the philosophy of mind that are of concern in this book. Selection enters the empirical discipline via the study of what has been called *focal attention*. From our perspective the study of attention should also be where we find clues about what the visual system picks out for further analysis. Clearly focal attention and what I have been calling Indexing are very similar. On the face of it the major difference would appear to be that we have several (perhaps 4 or 5) Indexes that work independently not only to select but also provide a reference to things in the world. Perhaps if we examine the experimental literature on visual attention we may find there some evidence about what sorts of things can be selected and also what the selection is for.

The general view in psychology is that attention is the mechanism by which cognition is able to make selections from among the various aspects of the perceived world, and that the ability to make selections is at the very core of how we interact with the world (Patrick Cavanagh refers to attention as the mechanism that “exports vision to the mind”, Cavanagh, 1999). This, however, leaves a great deal unsaid and raises questions that are at the heart of our present concern.

(1) The first question that the notion of selection raises is: *Why?* Why should the mind select and what role does selection play? The usual, and probably universally accepted answer, is that we must select simply because our capacity to take in information is limited. Being incapable of taking in everything, we must perforce select and we do so by applying what Donald Broadbent, one of the founders of modern information processing psychology, described as a *filter* (Broadbent, 1958). That the mind is limited and therefore has to be selective is unquestionably true, but far from being the whole story about the function of focal attention. (Even the part of the story that it correctly points to is highly incomplete. If the mind is limited, along what dimensions is it limited? And if it has to select, on what basis and along what dimensions does it – or can it –select?)

(2) It has also become clear that selection is needed not only to keep relatively unimportant or irrelevant information from clogging the mind, but it is also needed for reasons that have nothing to do with the mind’s having a limited capacity. It would be needed even if we were like the Martians in Heinlein’s cult science fiction novel *Strangers in a Strange Land* who could “grok” the entire perceptible world in one swallow. We would need it because in order to analyze and encode certain properties of the world we have to distinguish some parts of the visible scene from other parts; in particular, as Gestalt Psychologists pointed out in the last century, we must distinguish a focal figure from a background, or distinguish between a *this* and a *not-this*. Since perception makes such figure-ground distinctions for moving things as well as stationary ones, it implies that more than just selection is occurring, it implies that perception identifies the thing selected as an enduring individual independent of its instantaneous location. This, in turn suggests that there is a mechanism in perception that allows us refer to things in some way and keep track of their continuing identity. Thus focal attention may be thought of as a mechanism by which we pick out and refer to things we perceive (as Campbell, 2003, argued). FINST theory postulates a generalization of focal attention to multiple things (although FINST indexes



are different from focal attention in a number of important ways). As we saw earlier, we need to select several things at once in order to detect patterns among them. Thus the need to *individuate and refer to* things provides a second reason why we have to select items and why focal attention is a central concern in any discussion of how the mind connects with the world through perception. But there is yet another reason why we have to select certain sorts of things with attention – and indeed why what we select has to be things rather than places.

(3) The third reason we need selection has been explored both in the experimental psychology literature and in the philosophical literature. It is the fact that properties – or what finds expression as predicates – come in certain sorts of bundles or groups. The question of how our perceptual system manages to decode these bundles of properties has come to be called the *binding problem* (a term associated with the work of Anne Treisman, e.g., see the review in Treisman, 1988). When properties are properties of the same thing or the same sensory individual they must be marked somehow as *conjoined* properties, not merely as properties present in the scene. The earliest stages of vision cannot simply report the presence of properties. They must, in addition, provide some way to preserve the information that certain properties belong to the same sensory individual, so that we can distinguish between, say, a scene containing a green circle and a red square from a scene containing a red circle and a green square, or between a large striped animal coming towards us from a small striped animal going away from us, as well as all other combinations of these features. Indeed, the requirement holds across different modalities, so that the information must be in a form that enables us to distinguish between a green object that goes quack and a red object that goes moo. The problem of providing this information in the right sorts of bundles, which is called the *binding problem* (or which Jackson, 1997, called the "many properties problem" ), is crucial for our survival as well as being important in understanding how vision connects with the world. Although I reserve discussion of this problem until I examine Austen Clark's analysis in section 3.4, I mention it here because we will see that the solution involves object-based selection in a crucial manner.

### 3.1.1 Allocating and shifting attention: The role of objects and places

In recent years experimental psychologists have distinguished two ways in which attention can be allocated in a visual scene. One way, referred to as *exogenous* attention allocation, depends on events in the world – it is said to be *data-driven*. This form of attention allocation begins with an event in the visual scene that *captures* attention automatically without the intervention of volition or of cognition more generally. Some event – most notably the appearance of a new object in the visual field – captures attention (though other sorts of events, such as a sudden change in luminance, will do as well Franconeri, Hollingworth, & Simons, 2005). The other way of allocating attention occurs when you are searching for something and sweep your attention across a scene. It is called *endogenous* or voluntary attention allocation. An early demonstration of both types of attention switching is illustrated in Figure 3-1. Attention that has been allocated by these two means differs in a number of subtle but important ways. Exogenous or automatic attention allocation is the more important form of attention shift. It is more rapid and reaches a higher level of facilitation than attention that is moved voluntarily. Voluntary shifts of attention are easily disrupted by the automatic exogenous pull of visual events occurring at the same time (Mueller & Rabbitt, 1989; Rauschenberger, 2004).

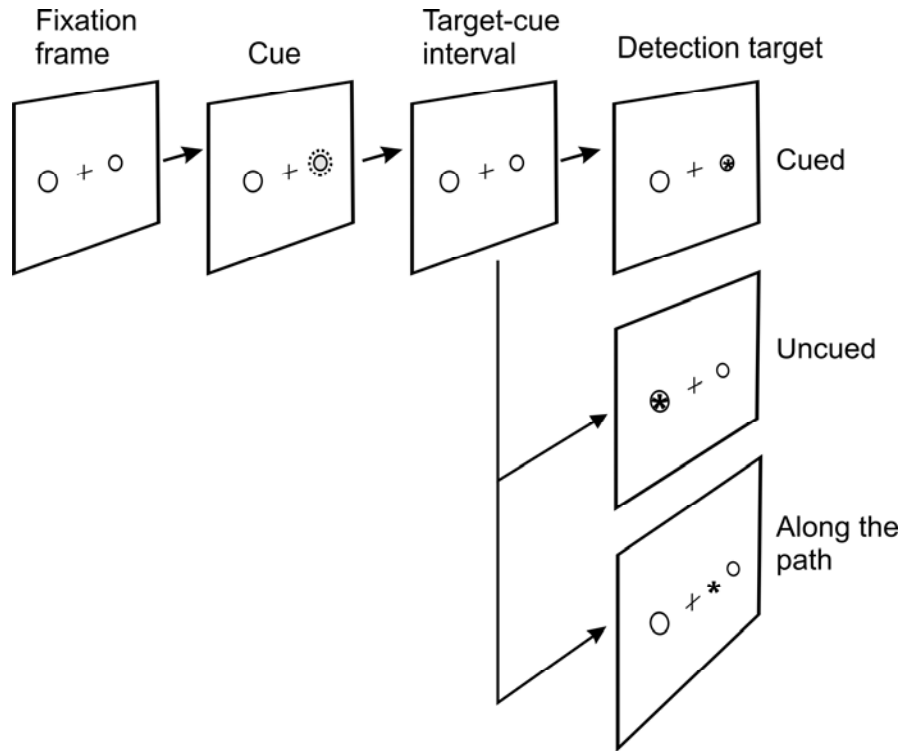


Figure 0-1. Illustration of how attention may be exogenously captured by a sudden brightening, shown in the second panel. Performance on a detection task (third panel) is better at the cued location than at the uncued location and at intermediate times is also better at a location along the path between fixation and where attention is moving to. For endogenous attention movement the subject is told (or shown by the appearance of an arrow at the fixation point) which direction to move attention. (Posner, 1980)

An important finding comes from a series of experiments and a mathematical analysis by (Sperling & Weichselgarter, 1995). These authors have made a persuasive argument that, at least in the case of automatically shifted attention, the locus of attention does not actually move continuously through space. The apparent movement may instead arise because of the distribution of activation over time and space when attention is captured and switches from one thing to another. According to this analysis, the apparent moving pattern may arise because the degree of attentional activation gradually decreases at its starting location and gradually increases at its target location. When these two spatiotemporal distributions are summed at intermediate locations it results in an apparently moving activation-maximum. Because voluntary shifts typically do not have a target event to which they are automatically drawn, it remains possible that they may sweep through intermediate positions. Although the data are not univocal on this question, it is plausible that when you move your attention voluntarily in a certain direction, you *may* sweep the locus of your visual attention through intermediate *empty* locations (so-called “analogue movement” of attention). I would caution against taking this as proven, however, since Jonathan Cohen and I showed that when people try to move their attention through an empty region in the dark (extrapolating the motion of a visible object that disappears behind an occluding surface) they probably do not move their focal attention through a continuous sequence of empty locations (Pylyshyn & Cohen, 1999). This conclusion is based on the finding that they perform poorly at continuously tracking with their attention the location where the invisible object is at any moment. Given a series of visible marks along the invisible path they do much better. Thus we concluded that their attention normally jumps from one visible feature to another using their highly precise *Time-to-Contact* estimation skill (for more on

the latter, see Tresillian, 1995). This conclusion is also supported by experiments by (Gilden, Blake, & Hurst, 1995) showing that when a subject tries to track the imagined continuous motion of an invisible object, the temporal pattern one gets in an adaptation experiment (when the imagined motion crosses an area adapted to motion) the transit times are consistent with the hypothesis that the moving attention consists of a series of static episodes rather than of a continuous movement<sup>16</sup> So movement of attention may not be continuous even in the case of endogenously or voluntarily generated movement of attention.

### 3.1.2 Attention selects and adheres to objects

A great deal of research in the past twenty years has convinced psychologists that viewing selective attention as being location-based is either incorrect or at the very least a secondary part of the story of attention allocation. An increasing number of studies have concluded that we attend to what I have been calling *things* (and what the psychological literature refers to as “objects”) rather than empty places. Evidence for this came initially from demonstrations of what is called *single-object advantage*. When a pair of judgments is made, the judgments are faster when they pertain to the same perceptual individual, even when all geometrical properties are controlled for, as shown for example by the experiment illustrated in Figure 3-2.

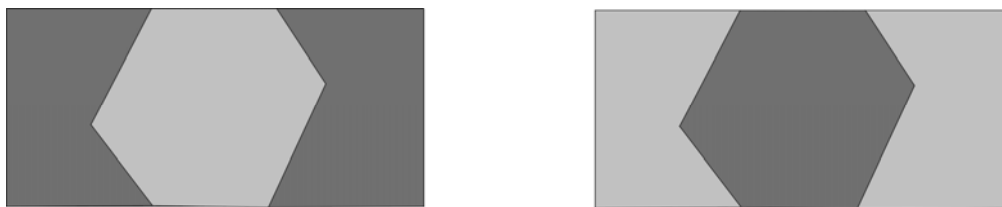


Figure 0-2. Task used in (Baylis & Driver, 1993) to illustrate single-object superiority. The task was to judge whether the left or the right vertex was higher. Judgments made when the vertexes were seen as part of a single figure were faster than when the vertexes were perceived as belonging to the same figure as opposed to two different figures. In subsequent studies (Baylis, 1994) the effect of other stimulus properties (such as convexity) was ruled out.

There is also evidence that the effect of attention tends to spread from where it is initially attracted to cover an entire larger object that subtends that initial attractor. For example, when attention is attracted to the highlighted end of a bar it then spreads to the entire bar. This spread does not simply terminate on an edge, but proceeds through what is perceived as the entire bar even if the entire bar is not explicitly indicated by a contour but is created by an illusory process called “amodal completion”, as show in Figure 3-3. Thus it is the bar as a perceptual whole object that determines how attention spreads. This suggests that the principles of attention-spread are modulated by the way that objects are parsed at some earlier or concurrent stage in vision. Thus the clustering operation mentioned earlier determine what constitute the units of attentional selection. Not surprisingly, they also correspond to the way objects are perceived in qualitative experience. This does not show that attention is directed to phenomenal objects (as

<sup>16</sup> What (Gilden, Blake, & Hurst, 1995) found is that imagined (i.e., attentive) motion in the direction of the adaptation was sped up while imagined motion in the opposite direction to the adaptation was slowed down. While this seems plausible, given the waterfall illusion, it is not what actually happens when a real perceived object moves through a motion-adapted region. With real moving elements the motion is always slowed down regardless of which direction the adapting motion had been in, presumably due to the fatigue of motion detecting neural circuits.

some have suggested), but rather it is the other way around: phenomenal objects follow the clustering pattern established by the parsing of a scene into objects by attentional selection.

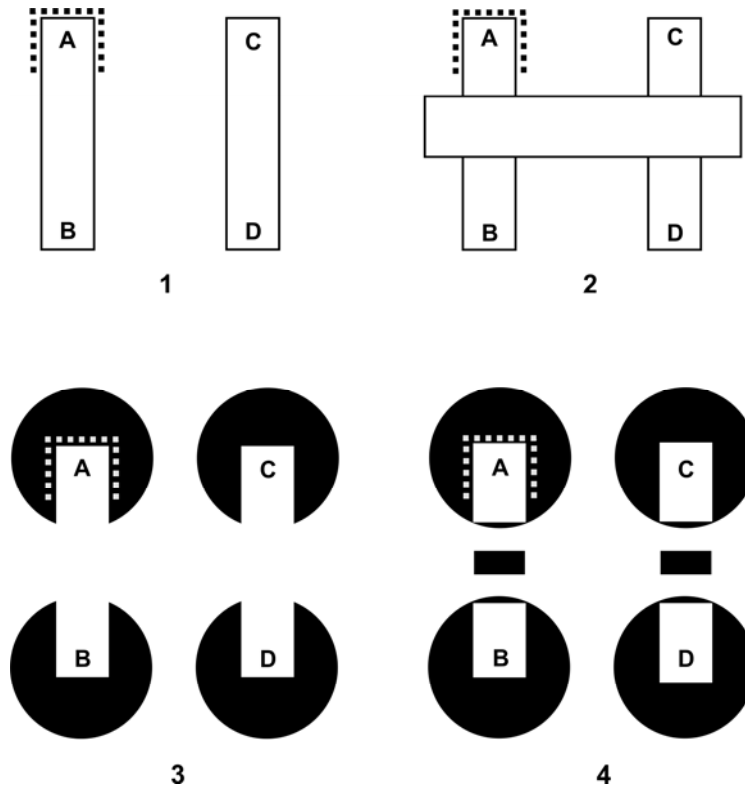


Figure 0-3. When attention is drawn to one end of a bar (marked A) by a cue (e.g., the brightening of the contour, indicated here by the dotted lines), its effect can be observed at the other end by the faster detection of a probe at that location (marked B), while the equally distant location on another bar (marked C) is not enhanced. This is true so long as A and B are perceived to be on the same bar (which they are in panels 1-3, but not in panel 4) (Adapted from Moore, Yantis, & Vaughan, 1998).

Even more relevant are studies showing that attention *moves* with objects being attended. A variety of phenomena of attention – including attentional enhancement (in the form of *priming*) and the negative effect of attention on the ability to re-attend to the same thing a short time later – show that attention appears to stick with objects rather than remaining at the attended location. The first case is illustrated by studies in (Kahneman, Treisman, & Gibbs, 1992) which show what is referred to as Object-Specific-Priming-Benefit (OSPB). In such studies the time to make a judgment about the identity of a pattern, such as a letter, is shortened when the letter occurs in the same object (typically a box frame) in which it had previously occurred (there is even evidence that such priming may last up to 8 seconds, Noles, Scholl, & Mitroff, 2005). A similar phenomenon has also been found with MOT displays, of the sort described in Chapter 2, where objects that are tagged by blinking show attentional facilitation even if no explicit tracking is required (Haladjian & Pylyshyn, 2006). The second case is illustrated by the phenomenon called *Inhibition of Return*, wherein attention is slow to return to something that was attended some 300 to 900 milliseconds earlier. Figure 3-4 illustrates an experiment showing that Inhibition of Return appears to move with the formerly attended object, rather than affecting the place that had been attended.

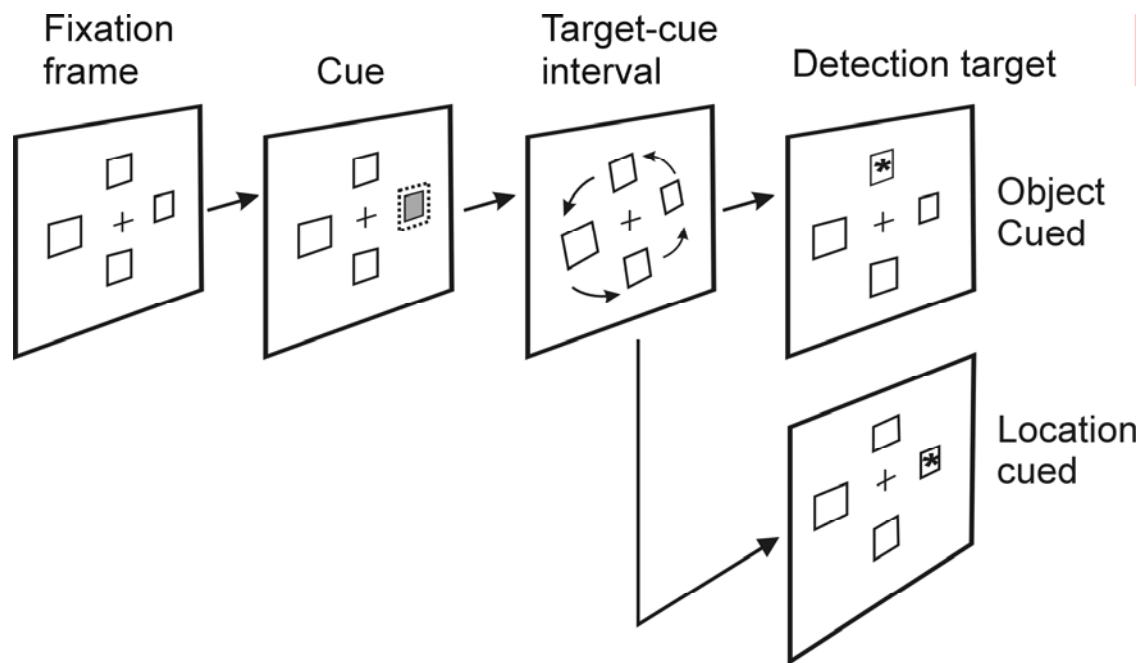


Figure 0-4. When attention is captured exogenously by an object and then disengaged, it takes more time to re-attend that object. This *Inhibition of Return* appears to move with the attended object (Tipper, Driver, & Weaver, 1991).

There has also been a suggestion in some of these studies that location may also be inhibited in this case (Tipper, Weaver, Jerreat, & Burak, 1994). But the locations in question in these dissenting findings are typically either nonempty or they are in a simple relation to indexed objects (e.g., halfway between two objects). It is doubtful that a location in a uniform unstructured field (what is called the *ganzeveld*, Avant, 1965) can be selected or inhibited – indeed after a few minutes of staring into a *ganzeveld* people tend to get disoriented and cannot even look back to some location they had just attended. Without objects in the field of view attention appears to be unable to get a grip on an empty location (failure to find inhibition at empty locations was also reported recently using Multiple Object Tracking, Pylyshyn, 2006).

Visual attention is a much-studied phenomenon and a great deal of evidence is available, of which I have presented only a few illustrative examples (a few more studies are described in Chapter 5). From our perspective what is important is that there is considerable evidence that sensory objects attract and maintain focal attention and that the evidence for the more common-sense notion that attention can be allocated to empty spaces is far from being univocal.

### 3.2 Selection and demonstrative reference: The role of FINSTs

In Chapters 1 and 2, I presented the outlines of a theory of visual indexing called the FINST theory. According to this theory, things in the world capture (or as we sometimes say, *grab*) one of a small number of available FINST Indexes, which thereafter are available for referring to the things whose properties were the cause of the capturing. In describing matters in this way I emphasize the role of an index as a *reference* mechanism. Indexes act like pointers in computer data structures: they provide a reference to some sensory individual (the nature of which has yet to be specified) without themselves serving as a code for any property of the individual that is indexed.

The parallel between pointers in a computer and FINST indexes is quite exact and helps to clarify what is being claimed, so it deserves a brief aside. The terms *pointer* or *address* are misleading in that they both connote a location. But in fact neither refers to an actual location. When we use the term pointer we are referring to a different level of abstractness from that of physical locations. Moreover, a pointer does not pick out a location in any kind of space, in even the most extended sense, including some sort of “functional space” since such a “space” does not have the minimal properties of a metric space – as I will argue in Chapter 5. A better way to view a pointer is as a name or a singular term. But the question still remains: what does such a singular term refer to? I claim that in the case of FINSTs it refers to what I have been calling *things* or *Fings*, meaning sensory individuals, or visual objects. Not *places where things are located*, but individual things themselves. At the end of this chapter I will return to the question that people often ask: Why don't I simply claim that they refer to physical objects? And why do I insist that the indexes do not select things by pointing to places where things are located? More on this presently.

### 3.2.1 Causes and codes

We have a strong intuition that when we detect a particular color, texture or shape in a visual scene, we detect it *as being at a particular location*. Indeed we feel that we can't detect a property without detecting its location. There are two questions at issue here that need to be kept separate. One is the question of which properties enter into the causal link between things in the world and the mechanism that does the primitive picking out – or in our terminology, which properties cause indexes to be grabbed. Another is the question of which properties are encoded (i.e., represented) and made available to the mind when an index has been assigned.<sup>17</sup> Another way to put this is in terms of the Kahneman & Treisman theory of Object Files, introduced briefly in the last chapter: we can ask which properties cause an object file to be created and which properties are encoded and entered into the file. These can be – and in fact very likely are – different properties. The distinction is reminiscent of Kripke's distinction (Kripke, 1980) between properties used in *fixing* the reference of a term, and the meaning of the term.

The puzzle of how some property can cause the capture of an index without itself being encoded may be exacerbated by the fact that in a modern computer it seems as though any signal that affects the operation of the computer serves as an input and therefore is encoded. But if you look in detail at how the input to a computer is processed you see a clear distinction between reacting and encoding. When you press a key on the keyboard two things actually happen. One is that there is an interrupt – a high priority signal is sent to the processor that tells it to stop what it is doing and attend to the interrupt (there are many possible types of interrupts – in Windows, the one from the keyboard is usually called IRQ1 or Interrupt Request #1). Getting an IRQ1 signal tells the computer that a keyboard event has occurred but it does not tell it which key was pressed (or released, since both cause interrupts). At that point nothing has actually been encoded but control is passed to an Interrupt Service Routine (ISR). Only then is the keyboard checked to see which key had been pressed. I mention this process because the two stages (interrupt and test) correspond respectively to the stage at which a FINST index is captured

---

<sup>17</sup> There is also a third distinct question that I will not discuss: Which properties are responsible for the object being tracked, i.e., for maintaining the index binding. Evidence suggests that properties such as color, size and shape are not used in tracking. If objects have distinct colors, shapes, etc tracking does not improve and changes in these properties during tracking are not even noticed (Bahrami, 2003; Scholl, Pylyshyn, & Franconeri, 1999). On the other hand, the speed of objects' motion, how close they are allowed to come to one another, and probably the rate at which they change direction does appear to affect ease of tracking, though they too do not appear to be encoded.

(grabbed) and the stage at which certain properties of the thing that caused the FINST assignment may be encoded and stored in its Object File (Section 2.3.1). These are independent stages and you can have one without the other. Moreover in a computer many events do not cause an interrupt and many interrupts do not result in anything being encoded. Another example of the distinction between index assignment and property encoding was discussed earlier (in Section 2.5) when I described the study by (Tremoulet, Leslie, & Hall, 2000) in which they showed that infants under 10 months of age could use a property (color) to individuate objects and form an expectation as to *how many* there are behind a screen and yet they do not use the same property to recognize *which* objects were behind the screen.

There are two kinds of properties that make certain information-carrying states representations or codes: (a) The first are *extrinsic*: these are the relational properties in virtue of which the codes are about something or other. It is through this semantic reaching-out character of codes that the organism is able to maintain a certain epistemic contact with the world – to refer to things, to select and track things and to bind the codes to things in the world. People have recognized two types of semantic relations in which codes participate; those that encode in terms of conceptual categories, making it possible for the codes to serve as constituents of thought, and those that, while they have representational content, do not function in rational thought. Codes of the second type are sometimes referred to as *nonconceptual*; they encode in terms of categories that do not function in thought. They may also represent in terms of categories, but the categories are not accessible to the rational cognitive system of the viewer. They may represent in what are sometimes called “subpersonal” modes. For example, they may represent properties of the proximal stimulus or properties useful for motor control or other actions (e.g., moving attention). I will have more to say about such nonconceptual codes in the next chapter (where I will propose that they are extremely limited – in fact that they consist only of the direct reference to objects made by FINST indexes). (b) The second type of properties that determine whether the information-carrying states are codes are *intrinsic*. They are properties in virtue of which the codes take part in computational processes. Such codes have to be instantiated by properties to which the system, qua computational process, is sensitive. In other words they have to be instantiated by properties that function to make the system work as a computer or a mind. Not every property to which a system reacts is a code, only those that affect its computational states in the relevant way – knocking the computer over or replacing one of the memory chips does not count as encoding information (for a more detailed discussion of this point see Pylyshyn, 1984).

### 3.2.2 Conceptual and nonconceptual representations and quasi-representations

#### (1) Conceptual representation

The term “conceptual” is used a great deal in discussions about mental representations, though its use has usually relied on the intuitive notion of its being “language-like”. A conceptual representation is one that is in principle expressible in language because it encodes the world in terms of individual concepts. Concepts, in turn, are very much like lexical items: They are individual symbols that represent categories and that take part as constituents in combinatorial systems that express thoughts, beliefs, goals and other typical cognitive states. This much is easy since they are the clear cases, recognizable as thoughts, beliefs and other “propositional attitudes”. But what about the encoding of such properties as those that are needed in theories of visual processing; properties of the proximal image, including edges (and

their properties such as spatial frequency – whether they are sharp or gradual, their orientation, polarity, type – concave, convex, occlusions, cracks, etc), properties of surfaces (such their depth profile and type – e.g., convex or concave – their texture, the 3D orientation of tangents and perpendiculars at points on their surface), values and patterns of lightness, shadows, velocity profiles, orientation of median axes in 3D, presence of geon types (which are the basic constituents of shapes, at least according to some theories, e.g. Biederman, 1987), and so on, as well as comparable properties in the auditory modality (temporal patterns of fundamental, harmonic and aharmonic frequencies, apparent location of sound sources, phonetic content, and so on).

The proprioceptive modality probably offers the greatest array of properties that we use every minute of our waking day but that we cannot describe, both for lack of terms and because we are not conscious of them. There are countless proprioceptive signals that allow us to maintain our posture and balance, to walk upright, to avoid obstacles while moving, to reach for things and carry on all our mechanical commerce with the world. These are generally not conscious – in fact in most cases they do not produce any noticeable sensations (do you sense the contractions and expansions of your muscles as you sit without slumping?). Some, though certainly not all of these sorts of sensory inputs may be represented – stored for future use or processed to predict the future state of the world or of your body or to plan a sequence of actions. Whatever you think that concepts are, much of this information is not conceptual. Not only do we not have terms in our language for them, but we could not express them in language because, among other things, we are not aware of them at the personal level. These are properties which, if they are represented at all, are represented sub-personally in terms that have meaning only to subpersonal computational mechanisms and not to whole persons. And not every sensory property is represented. Some sensory information is represented and some not; some is represented in terms of categories that could potentially be available to cognitive processes (we could in principle reason about them) and some not. For example if you are a linguist you might reason about phonemes and NPs or VPs, though not in the same way that these sub-personal concepts are processed within the language parsing system. The boundaries between nonconceptual and conceptual on the one hand and between conceptual-personal and conceptual-subpersonal on the other hand are all distinctions that we can appreciate from the clear cases but for which we cannot provide operational criteria. The same is true of the distinction between representations whose contents are consciously available and those whose contents are not available to introspection. I will have more to say about that distinction in the next chapter.

## *(2) Nonconceptual representations and non-represented properties*

In recent years it has become clear that there is a great deal of information of which we are not aware, that serves purposes other than object recognition and belief fixation. Among the information that is causally efficacious is information relating to the locations of features and their distances relative both to other features and to the organism itself (by “feature” people usually mean any sensory property, but more typically a basic spatially-local property). This information plays a role in defining feature-clusters and other Gestalt-like groups, and in pairing space-time features to create the perception of apparent motion. There is also clear evidence that such information can serve with remarkable precision in shaping our actions, both voluntary and involuntary (e.g., it allows you to keep your balance while walking).<sup>18</sup> It is also clear that

---

<sup>18</sup> A remarkable example of how much we rely on unconscious proprioceptive information in walking and maintaining our posture is provided by cases of patients who are essentially deafferented and do not have proprioceptive input. They can, with great difficulty, replace the proprioceptive information with visual information but only with extremely arduous training and persistence. See the description of one such patient in (Cole, 1995).



information about *location* plays a central role in many such unconscious processes. Does this entail that the location of features is *represented*, as claimed by many philosophers who speak of them as constituting a form of nonconceptual representation? Austen Clark (Clark, 2004) claims that it does; that perceptual grouping of features on the basis of their proximity to one another, as occurs when we construct feature-clusters, requires encoding the location of the features in question.

The notion of nonconceptual representation was introduced largely to acknowledge not only that there is information that is unlikely to be conceptualized (since, for example, we are not aware of it, or it never enters directly into beliefs or other cognitive states), but also to accommodate information that is extremely fine grained and that appears to be functional in at least the sense that it includes discriminable sensory differences, yet is very unlikely to be conceptual (or even conceptualizable). The category of nonconceptual representation covers many different kinds of cases (see, for example, the essays reproduced in, Gunther, 2003). This includes information that clearly is relevant to our actions (such as reaching and grasping or walking) yet is below the radar of our awareness and our rational thoughts. Moreover, it often involves such precise magnitudes that it does not seem as though we could in principle conceptualize and reason about it (for example, when we hit a golf ball, or throw a basketball into a net or play a piece on the piano). There are representations that we do not want to call conceptual only because they do not involve the sorts of categories that people are aware of or could be made aware of. One example of this are representations computed in the course of a visual analysis – that involve, for example, such categories as edges, boundaries between light and shadow, orientations of edges, types of textures, surface orientations at a point, and other properties of the retinal stimulus that we as perceivers do not reason about, but which the encapsulated vision module has to compute over (Fodor, 1983; Pylyshyn, 1999). These are the sorts of categories that show up as labels in Marr’s “full primal sketch” and “2 ½ D” representation. They are in every respect like conceptual representations except the categories are sub-personal and are not available outside some modular process.

The most widely cited evidence in favor of the assumption that there is a type of representation that is nonconceptual appeal to the apparent differences between the contents of conscious experience and the contents of thoughts.<sup>19</sup> This manifests itself in at least three ways: *First* it is clear that we can consciously distinguish among finer differences (e.g., finer gradations in hue) than we could reasonably assume we have concepts for (assuming that concepts are built from a finite, and reasonably small base that is used combinatorially to build complex concepts, much the way languages do). *Second*, there is a degree of independence between a proximal stimulus (i.e., retinal pattern) and a percept, and this needs to be explained. Take, for example, the way an ambiguous figure such as the Necker cube (shown in Figure 4-1), changes between distinct appearances even though both the proximal stimulus and what we believe about the figure remains the same. What is it that changes? According to one view what changes is how we represent the figure in nonconceptual form. Because this way of understanding nonconceptual representation rests on certain views about the contents of our conscious experience, I will leave an expansion of this topic for the next chapter. *Third*, there is some reason to think that the content of experience – the full fine-detailed information that we experience, is available for a very short time before it either fades or is encoded conceptually. The idea that detailed nonconceptual information is available briefly is consistent with many findings in the experimental literature that provide evidence for what is often called “iconic

---

<sup>19</sup> For more on the role of conscious contents in relation to forms of representation, see Chapter 4 (e.g., section 4.1), as well as the discussion of representation of space in section 4.4 and Chapter 5.

storage”. For example, (Sperling, 1960) found that a great deal of visual information is available for a fraction of a second after the information input (i.e., the light) has disappeared. This so-called iconic store does not contain conceptualized information (in fact it is commonly referred to as pre-categorical store). Such information is available for any visual process that itself does not take more than a few hundred milliseconds – including processes that encode information conceptually. While this third option does constitute a type of nonconceptual “representation”, calling it a representation may be misleading insofar as it need not involve any encoding, but may arise merely from inertia or hysteresis on the part of sensors. It may lack the signature properties of representations: it is arguably just a geometrical projection of retinal stimulation (at least according to one view, championed by Sperling) so it cannot misrepresent and its content (what it represents) does not enter into any generalizations (see (c) below for more on these criteria for being a representation).

My concern at this point is primarily with the question of whether certain spatial properties such as location and distance are encoded (represented) early in vision and whether they can serve as the basis for primitive selection. For this purpose we should recall that there is a substantial difference between the claim that certain properties, such as the locations and distances between objects in the world, play a *causal* role in a visual process, and the claim that these properties play a role by virtue of being *represented*. Objects are always at some location or other, and the effect they have on a perceiver may depend on where they are, but the locations they are at (even their locations 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 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 result of such mappings as “neural layouts” or NLs)? Such neural layouts help to illustrate the general theme that there are many types of representations, ranging from conceptual, through subpersonal to informational states that are scarcely worth calling representations at all. 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 map-like representation (I will have more to say about map-like 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 a 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 anyone is free to call it a representation if they wish, on the grounds that NLs usually carry information about something in the world to which it is causally connected. 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 between states whose representational *content* plays a role in explanations and those in which the content (if any) does not. If we gain no explanatory advantage by specifying *what* a NL represents, then it is not relevant to the question of explaining how mental representations connect with the represented world. The fact that the NL may look 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 a number of 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, directions, in the NL correspond to places, etc in the world but also that they determine the organism's behavior vis-à-vis those represented places. In other words, 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. So far as the principles go (at least as understood by people like Kohler and Wertheimer) they are principles expressed over properties of the proximal stimulus or even over neural fields in the brain,<sup>20</sup> not over locations, distances and directions in the world.

A 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 a NL in which frame of reference it represents an object's location, since by itself it does not represent an object as located anywhere. But in the case of representing in the strong sense, where the NL functions to direct movements or to identify objects, it *does* matter how its spatial relations are represented. In that case a 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

---

<sup>20</sup> Processes operating over NLs typically respond to spatially-local properties of the NL – they operate over “local support” (see note 14). The principles of operation of such processes are prima facie expressible over nonconceptual neural properties. There has been a recent 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 and derive Gestalt clusters, solve the correspondence problem in certain cases, and even carry out tracking (examples are found in Koch & Ullman, 1985; Pylyshyn, 2003, Appendix 5A). Thus they may be appropriate for the sort of neural field processes envisioned by Wolfgang Kohler (Kohler, 1947).

<sup>21</sup> I am leaving out a lot here. For example, what makes it the case that a typical terrestrial or road map could misrepresent is that maps generally are constructed by a map maker with the intention that certain features of the map 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 thus for the map to misrepresent. These degrees of freedom are absent in the case of NLs unless we assume that the map is interpreted by some process in such a way that there is 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 relevant mechanisms – and (Marr, 1982) motivated his analysis by asking what various visual mechanisms were *for*. Fred Dretske (Dretske, 1981) suggests another way in which an information-carrying state might misrepresent – a way based on learning. If the system had been exposed to examples of property-instance pairs it could learn which features of the environment the information represents and thus could be in a position to misrepresent these features. These are all questions that I will not get into, beyond making the point.

actions *it matters how the location is represented* (or what it is represented as). Without this sort of strong sense representation, with only a direct object-to-NL mapping, there is no possibility of misrepresentation and it is misleading to call the NL a representation or a map.<sup>21</sup>

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 the existence of a NL means that these layouts are spatial representations – 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 on NL or properties in the world? Suppose, for the sake of argument, that the clustering algorithm applies only over distances on the neural layout which, in turn, is a homeomorphic transformation of activity on the retina.<sup>22</sup> 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 2D distance on the retina, a 2D distance far away from the observer, or a distance in 3D which is 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.

(4) *When should we postulate representations?*<sup>23</sup>

The purpose of postulating representations in general 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 case of 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

<sup>22</sup> These examples are for purposes of illustration – I am not at this stage 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 neural layout. If they applied to distal properties then the present argument would not work, but then again neither could we claim that the neural layout was the basis for the clustering or correspondence solution since from what we know about V1 (or any other neural layout) it is prior to processes that provide 3D information (prior to the constancies). There have been conflicting claims on the question of whether 3D properties are relevant to apparent motion, some investigators maintain that 3D distance is relevant (Attneave & Block, 1973; Wright, Dawson, & Pylyshyn, 1987) and some that it is not (Ullman, 1979). In recent years there have been many reports of 3D properties being relevant to what seem like early processes, such as popouts in search (Enns & Rensink, 1990; Rensink & 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 & Franconeri, 2006; Liu, Austen, Booth, Fisher, Argue et al., 2005). These suggest that such processes are post-constancy or post-depth analysis, and therefore that they do not involve (only) places on the neural layout (in V1).

<sup>23</sup> Michael Devitt (Devitt, 2006) recently coined the term “Pylyshyn’s Razor” to refer to the idea that we should endeavor to minimize the power of the explanatory mechanisms to which we appeal, and that therefore one should not postulate representations where more restricted forms of architecture-based (non-representational) mechanisms would do. I have argued for such a principle in various places (Pylyshyn, 1984, 1991, 1994, 1996).

states involved. In discussing the way information might be carried by a neural layout, I noted earlier that an explanation might sometimes be stated in terms of the physical (or connectivity) distance between places on the layout (or individual units of a neural network) (indeed such networks have been proposed, including one in the Appendix of Chapter 5 of my book, Pylyshyn, 2003), in which case nothing is gained by treating such clustering processes as operating over codes or representations since nothing hangs on whether or not you take such distances as representing distances in the world.

The above discussion reflects a general preference towards a conservative use of representations in theories. Recently I have argued that we should not postulate representations if no explanatory advantage is gained by such a postulate. In (Pylyshyn, 1991) I suggested that in general the preferred explanation is one that relies on assuming the weakest mechanisms compatible with the evidence – in other words on assumptions about mechanisms that are constrained in the patterns of behavior that they are able to accommodate. There is nothing unusual in this idea. Given a body of evidence we generally prefer the explanation with the fewest degrees of freedom or the fewest *ad hoc* assumptions. But considering all the relevant evidence we have, our account will always require *some* assumptions or principles postulated to account for the evidence at hand – those are the degrees of freedom we have for taking account of the evidence in giving explanations and in predicting future observations. There is, of course, always the problem of individuating and counting assumptions. But however we do that, we give extra weight to the least powerful set of assumptions or, in the case of a computational theory, to the weakest mechanism compatible with the observed behavior. Explanations based on representations are the most powerful – with them we can construct a Universal Turing machine that can emulate any mechanism.<sup>24</sup> So we prefer to appeal to the weakest mechanism, which is one based on fixed structures, as opposed to the highly malleable representations. This leads to the principle that we try to account for as much as we can based on what I and others have called the architecture of the system (Pylyshyn, 1980) before postulating representations.<sup>25</sup> An explanation based on properties of the architecture not only appeals to a more constrained mechanism, but unlike the more powerful representation-based explanations, it also makes a specific commitment about what functions one should find in the neural substrate. Of course there is a tradeoff; the weaker the power of the postulated mechanism, the less likely it is that it will be adequate to account for the evidence as more and more evidence is accumulated. This is how in cognitive psychology one is forced sooner or later to attribute more and more functions to representation-based mechanisms.

---

<sup>24</sup> It has not always been recognized that the Universal Turing Machine (UTM) is just another TM. What makes it a UTM, rather than the *n*'th TM is the way we view the symbols on its tape. A UTM only comes into being when we (the observers) interpret some of the symbols on its tape as representations – in particular as representing the identity of or the program for another TM, the one whose behavior the UTM is simulating.

<sup>25</sup> This is also closely related to other razors. De Morgan's Cannon says that one should not postulate higher level psychological functions, or functions associated with organisms higher on the phylogenetic scale, when lower psychological functions would fit the evidence. It is sometimes considered to be a special case of Occam's razor which is a general principle of preferring simplicity in theories, where simplicity is an essential although undefined commodity.

### 3.3 Problems with selection-by-location

I return now to the topic with which I began this chapter, the question of whether we select places or objects. Notwithstanding the evidence for object-based attention, both psychologists and philosophers tend to view location as the primary property in terms of which selection occurs, and therefore as the property through which the mind addresses things in the world. This view is consistent with the informal observation that if you want to look at something or touch something you need to know where it is. There have also been a number of studies (reviewed in Pashler, 1998) showing that in most cases where an object is correctly identified, its location can also be reported correctly, from which many investigators conclude (e.g., Pashler, 1998, p97-99) that location is the basic property used for recovering other properties. For example, there is evidence that the way one finds a combination of color and shape (i.e., when searching for a conjunction of these two properties) is by using one of these features (e.g., color) to find its location and then to use its location to test for the second conjunct (e.g., shape) at that location (Pashler, 1998). Mary-Jo Nissen (Nissen, 1985) tested this view directly by examining the probabilities of various conjoint combinations being found when searching for conjunctions of properties. She showed that in searching for a target with a specified color and shape the probability of finding the right color and of finding the right shape were not independent, suggesting at least that there was some way these two interacted in the search for their conjunction. Nissen then showed that the probability of finding a given color and the conditional probability of finding a particular shape given that attention is focused on its location were *independent*, lending support to the hypothesized two-stage process (e.g., to determine whether the conjunction of properties  $P_1$  and  $P_2$  is present, (1) find property  $P_1$ , get location  $L_1$  of  $P_1$ , (2) switch attention to  $L_1$  and check whether  $P_2$  occurs there). Of course in all these cases, where objects do not move, being at a particular location and being on a particular object are perfectly confounded. Moreover, it is quite reasonable that priority may be given to location *once the object has been selected*. There is considerable evidence for the claim that location is treated as special among the various *conceptualized* properties of selected objects. Indeed, our work on multiple object tracking shows that when objects disappear, the location of targets is retained better than any other property of the targets (Scholl, Pylyshyn, & Franconeri, 1999). We also showed that the disappearance of tracked objects for half a second, and perhaps even much longer, does not disrupt tracking if the objects reappear where they had been when they disappeared, thus showing that the location of disappearance must have been retained in some form (Keane & Pylyshyn, 2006). But our assumption is that the disappearance itself causes locations to be conceptualized and stored in memory. It's not clear whether in this case the relevant memory is the object file that was associated with that object at the time of its disappearance, or whether it uses some other sort of memory such as Visual Short-term Memory. If it's the former it would suggest that an object file can exist, at least for a short time, even if it is not connected to an object.

#### 3.3.1 A note on the role of location in selection and tracking

The proposal that we select and track visual objects based on the continuing individuality of the objects (what philosophers call their "numerical identity") has sometimes been looked upon with suspicion, as though what was being proposed is some sort of magic wherein an object is mysteriously picked out and tracked without any physical connection. It seems to me that there are two reasons for this suspicion. One is that every object is in fact at *some* location and when one selects it one is selecting an object that happens to be at a particular location. Thus there is the usual temptation towards the intentional fallacy here as in other cases of representation

(where properties of the referent of a representation are mistaken for properties of the representation itself – a type of use-mention error). In fact not only is an object at a location when selected, but it may even be selected *because* of its location, since one clue that there are several individual objects is that they are at several locations. As I pointed out in Section 3.2.1, many properties of objects may contribute causally to the object being selected. But that does not entail that any of those properties are encoded or represented nor that these representations then play a role in keeping track of the identity of individual objects. The claim that a certain property may cause a particular item to be selected (or that may grab an index), yet not be encoded is also a claim that some find puzzling.

When an index is assigned to some individual, the process presumably involves a causal chain that eventually causes sensors to respond to a local region of the proximal (e.g., retinal) stimulus. Such processes respond to stimulation at particular location(s) and are sensitive to at least a local distance metric since one of the things they presumably do is compute a cluster of local properties based in part on their proximity to one another. But it is a separate question whether properties such as the size of the cluster or its retinal location are *represented*. It should be clear that there is no requirement that such metrical information be *encoded* and entered into an object file in order for an index to be captured or an object file to be created. Whether it needs to be represented in some other sense is a separate question I will consider next, and to which I will return in chapter 5 when I consider the issue of how spatial properties might be represented.

Some people are suspicious about the object-selection principle because they have trouble imagining how information about the properties of an object can be encoded without knowing the objects' location, and how an object can be tracked without keeping track of where it is. Before looking closely at the MOT case let's consider this question somewhat more broadly. There clearly must be a causal information stream coming to and/or from the object for a FINST to be assigned or for the object to be tracked. A FINST assignment often occurs when a new object comes into view, but it can also occur when there is some other sort of transient (e.g., the object briefly becomes brighter). Can a FINST index be assigned if the location of the transient is not known? If you think about the various current technologies that select things, you see that there are many ways that selection can occur, other than by using information about the location of the selected object. I have already given the example of a pointer in a computer which selects (and retrieves) a symbol that is the referent of a particular given symbol, say "x" (where in computer science the such a symbol is may be referred to as a name or a variable or a pointer). It does this without knowing where the referent of "x" is (and indeed, in a modern multi-tasking computer there is no specific place where the referent is located for more than a few milliseconds at a time). It is able to do this because of the architecture of the computer, which means that the explanation must be given at a different level or in different – noncomputational – terms (e.g, electromechanical in the case of a computer, neurophysiological in the case of the brain).

Other ways of selecting things requires identifying information to be transferred between the "x" and its referent and/or the other way around. This is the case with cell phones that find the set you are calling without knowing where it is. Other selection methods use different ways of identifying the referent of "x"; a tuning fork can be found by emitting its specific tone, a piece of metal can be found by using a powerful magnet (as people who have gone into an MRI room with keys have discovered). None of these use location as *the way of selecting* its target, so not using location should not be too mysterious. What is essential for accessing a target is some way of identifying it, which is why it is essential that the FINST mechanism track the relevant object and maintain its identity from the time it was visually distinct (at the start of an MOT trial).

The minimum function needed for an object to have the right kind of causal or informational link with a symbol token is that there be some causal or nomologically-supported *dependency* between the object and its associated symbol token (this is similar to the informational view of reference, as developed by Dretske, 1981). What kind of dependency? Any simple causal connection will do. One simple candidate is that if an object appears in the field of view (under the right background conditions of lighting, size, location relative to the fovea, speed of motion, and so on) a symbol token appears in the appropriate part of the early vision system. To a first approximation, a symbol might be tokened in the early vision module while, and only while, the object is in view (it's an approximation because one needs to deal with the persistence of the symbol over some reasonable time after the object fades – a qualification that needs some spelling out). The symbol is unique to that object token on that particular occasion only. If the object moves around, nothing changes in the symbol token – whatever relation it had with the object, such as the above dependency relation, remains. Nothing about such a dependency mentions location (except in the general background conditions). Such a dependency relation establishes a causally-supported symbol-object pairing which is an important beginning. But is clearly not enough for the purpose of the FINST theory. The FINST proposal requires that once a dependency relation is established, an information channel opens up so that, for example, some operation performed on the symbol token has consequences that depend on properties of the object. For example, if the symbol is in the appropriate relation to the system architecture (e.g., if it is placed in a certain register) some other designated register would change contents depending on the color (or shape or any other sensory property) of the associated object (a possible neural net implementation of a similar function is provided in Pylyshyn, 2003, Appendix 5A). This, in effect, has the consequence of allowing a function to be executed that returns a property of the relevant object, which is another way of saying that the index can be used to bind the argument of a function or predicate to the object in question. Mental functions don't normally reach out and channel information to things in the world, but only to representations. Yet once the right kind of dependency relation holds between an object and a symbol token (once an index has been established) it can, in effect, be deployed to query the object. Whether this way of putting it is the most perspicuous remains to be seen. But the idea is that some such minimal functionality is necessary in order for the information link between object and symbol token to serve the sort of function postulated in this discussion.

In selecting objects in situations such as those discussed earlier, including in MOT, the basic requirement is that there be a channel between symbol and object – a way for some part of the visual system to “communicate” with the particular object (e.g., send a signal to it, or test it for certain properties). Which part of the visual system is involved and the detailed story on how it “communicates” with the particular object are questions that likely to have only a neural-physical-optical account, which is to say that it may well be part of the fixed architecture. Indeed, as noted earlier, this is the sort of function for which neural networks of the connectionist sort are particularly suited. In fact the proposals mentioned earlier (Koch & Ullman, 1985; Pylyshyn, 2003, Appendix 5A) do provide a way of sending a signal to certain salient objects. These networks are interesting mainly because they show that a signal can be sent to “the most active neural cluster” without ever specifying *where* that cluster is – at least not in a way that could be used outside the neural layout. If we accept the view presented in the previous section, that NL's need not be representations in the strong sense, then this shows that a strong sense of representation of location is not needed in order to establish an access channel between symbol tokens in early vision and selected objects.

Consider how the selection processes work in the case of experiments motivated by the FINST hypothesis. Data reported in (Burkell & Pylyshyn, 1997) show that in selecting a subset



of items to search through, distance between items does not affect search, and data on subitizing (Trick & Pylyshyn, 1994b) shows that precuing the location of items to be enumerated does not affect subitizing, which we interpreted as suggesting that accessing Indexed (FINSTed) objects in search or enumeration tasks does not require having to locate them first (though with larger numbers of objects it does). In addition, data on MOT suggests that the tracking system either does not have location (or direction) information or, if it does, it does not (or cannot) use it with any precision in tracking (for example Franconeri, Pylyshyn, & Scholl, 2006; Keane & Pylyshyn, 2006, showed that when objects disappear behind an opaque occluding surface and reappear shortly after, they are not tracked better if they reappear at predicted locations going in predicted directions). The tracking process is assumed to use indexes to bind the target objects to some internal symbols or Object Files. This includes binding a target to the argument of the process that can switch attention to it. To execute  $\text{MoveAttention}(x)$  the hardware that carries out the command needs to be able to actually bind  $x$  to the object and thus to move attention to it. Just as we don't know in detail how attention is moved, so we do not know in detail how the binding occurs. If we can think of attention allocation as moving a "spotlight", as many people do, so we can think of tracking as involving a causal chain centered on the target. In fact there are a number of proposals for how an attention beam can be kept centered on a moving object (e.g., Kazanovich & Borisyuk, 2006, propose a neural model). The problem is well studied in computer science with both hardware (Gul & Atherton, 1989) and software implementations (Shapiro, 1995). Most of these involve clustering or edge-finding process which use information about relative location of features (typically based on a Neural Layout), but this information is only used locally and thus need not be available outside an encapsulated module, the way that routines in some programming languages have local variables whose values are unavailable outside the running routine.

Finally, just as we need to distinguish between interrupts and tests, so we need to distinguish between memory (which stores encoded information) and a mechanism's inertia or hysteresis or decay time. One of the MOT findings is that objects can be tracked even when they disappear for a short time (a fraction of a second). We can think of this as requiring memory, and therefore of storing an encoding of object properties (such as location, speed and direction of travel) or we can think of it as just being within the time constant of the mechanism. After all, we do not find it a theoretical challenge that people can track moving objects in MOT despite the fact that the screen is refreshed every 17 or 34 ms, so why is tracking over longer periods of invisibility a problem? The tracking mechanism may be complex and it may have various constraints built in, just as the rest of the visual system has such built-in constraints. For example among the empirical findings concerning tracking briefly disappearing targets (Scholl & Pylyshyn, 1999) is that when a target goes behind an occluding surface it is tracked better if there are occlusion and disocclusion cues as there would ordinarily be if the target moved behind an opaque surface and reappearing on the other side. So clearly even though the tracking mechanism computes simple functions that do not require access to general knowledge, there are natural constraints built in that determine the conditions under which it tracks best. In these we see a role for some local spatial and temporal integration (an apparent "quasi-memory"), although in every case what is involved is short-term or temporally-local and spatially-local computations, so it is compatible with the view that indexing does not require an encoded memory for location that is used to enable tracking.

More important theoretical implications of whether selection is by location, by property, or by individual object emerge in connection with a problem that I have already mentioned, known as the *binding problem* or the *many-properties problem* (the term used by Jackson, 1997), which I take up next.

### 3.3.2 Selection and the binding problem

Sensory properties come in certain kinds of bundles and one of the challenges of early vision is to make information about this bundling available for subsequent conceptual encoding. As I mentioned at the beginning of this chapter, this is one of the main functions served by *attentional selection*. The bundling I speak of is simply the fact that properties belong to things and it is important that the combination of properties that belong to the same thing be somehow indicated. What the earliest stages of vision must do in order to permit the subsequent conceptual description of a scene, is to present the information in such a way that properties that belong together, or are associated with the same tokens in the scene, are somehow flagged as such. Many mechanisms that have been proposed as part of early vision fail to do so. Consider, for example, the early “Pandemonium” model of (Selfridge, 1959), or its modern “connectionist” descendants, or the purely hierarchical views of initial encodings in which simple property detectors (simple cells) send activity to complex cells, which send activity to hypercomplex cells, until (at least in the minds of some theorists) ultimately something like the “grandmother cell” was activated, leading to recognition.

These proposals detect properties independently and leave it to subsequent stages to put them together into object descriptions. But it is not enough to indicate, for example, that certain properties such as redness or greenness or triangularity or squareness are present in the scene, because then higher centers of conceptual vision would not be able to tell that, say, the red goes with the triangle and the green goes with the square, or vice versa. The problem that this raises turns out to be completely general and rather deep, and is discussed in various forms by Strawson, Quine, Kaplan, and others. How does early vision circumscribe and report the information in a display such as in Figure 3-5 so as to preserve the information that in the first two frames it is the large square that has black dots and that the circle and small square are finely-textured gray?

The answer given universally to this puzzle, both in psychology and in philosophy is that vision represents properties, such as shapes and textures by their location in property-specific neural layouts called “feature maps”. The story is that it is the fact that *dotted* on the texture map is at the same location as *square* on the shape map and as *large* on the size map that puts them together – and it is this that allows vision to distinguish the three figures in Figure 3-5.

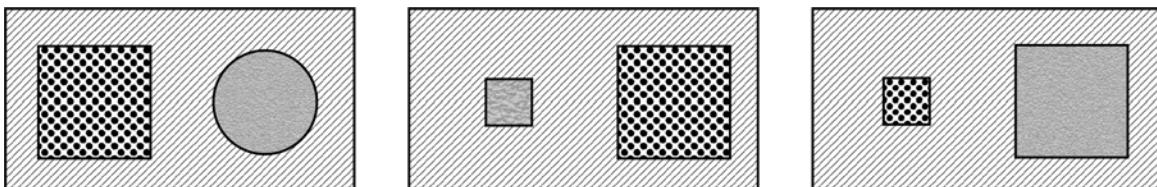


Figure 0-5. The early visual system must report information in a form that makes it possible to discriminate these figures composed of identical sets of shape, texture, size properties in different combinations. Specifying how it does so is called the binding problem.

In order to put the features together (Treisman & Gelade, 1980) postulate an additional ‘master’ map that allows the locations on the feature maps to be kept in registry and that allows focal attention to be moved across the scene independently of features, but in registry with feature maps. There is empirical evidence that discriminating objects on the basis of a conjunction of properties may require the application of focal attention. The story about how this happens can be told in terms of Feature Integration Theory (Treisman & Gelade, 1980).

According to this theory, in order to find the small dotted square among the three figures shown here, one makes use of two feature maps and a master map – all of which are part of the early vision system. Figure 3-6 shows the Feature Integration Theory applied to a group of colored shapes (shown in the figure as different textures). The stimulus in this example is registered on three maps. The *shape map* shows that there is a rectangle at one location, the *color (or texture) map* shows locations of the different textures and the orientation map shows the location of the different orientations. Each of these maps is coordinated with a master map that does not show the features, but does provide a way to control the search through the feature maps. So to locate the large dark horizontal ellipse, one checks the *shape map* for an ellipse. Then one checks the color or texture map for the texture. To solve the binding problem one simply finds the place where both matches occurred. One does this by either by using the master map to guide focal attention to places and then checks the properties corresponding to those places at the distinct feature maps, or by finding one of the conjuncts and checking its location for a feature on another feature map.

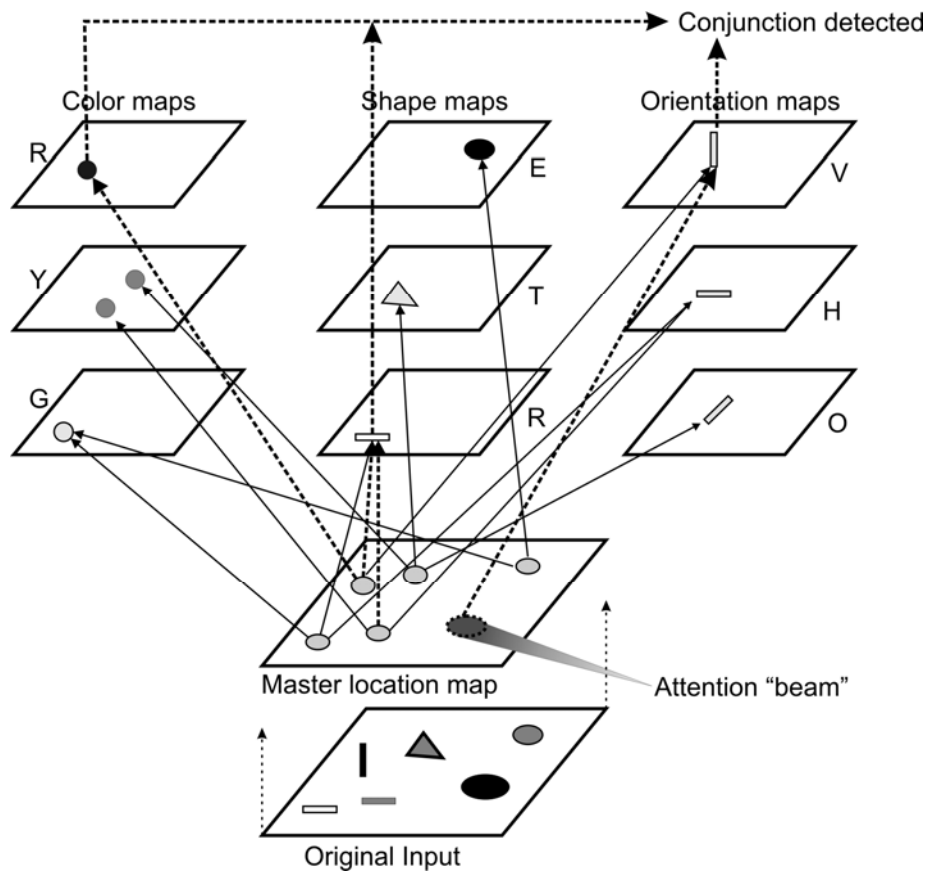


Figure 0-6. The Feature Integration Theory assumes a master map for coordinating and conjoining properties by their spatial co-occurrence (after Treisman & Gelade, 1980). A conjunction is deemed to occur when a property is detected and marked on its particular map and then attention confirms that the second property exists at the same location.

But there is a problem with this analysis, and indeed with any analysis that does not recognize the central role played by objects. To see why this is so, consider how to encode, say, the large square in Figure 3-5 as being at some particular location on the shape map. One is immediately faced with the problem of how to locate that shape so that it provides the appropriate cross-reference for locating a color. The square is not at a precise punctate location so one has to locate it at a region (Clark, 2000, does recognize the role of regions in a way that

FIT does not, though it does not help resolve the problem discussed here). If a fixed-size region is chosen then if the region is too small, it will not distinguish the small square from the large one in Figure 3-5. If it is chosen too large, it will not distinguish between the square and the circle, and if it does, it will not be in a position to determine what texture that region has since it will include both the texture inside and the texture outside for at least some of the shapes. So the region has to be just right in both size and shape – a requirement that assumes that the object in question has already been selected! The problem is that there is no precise place where squareness is located, and without a precise place for squareness there is no place to look for texture on the texture map. There is even evidence that what counts as the border for purposes of attentional selection depends on the perceived object – the same border may sometimes serve as the boundary of selection and sometimes not (see Figure 3-3). This problem will not be solved by an easy fix, such as allowing the region to fill out to some border, because what counts as the relevant border is the *outside edge of the object* whose shape is represented, or by letting location be a probability distribution, as Clark suggests in his more recent paper (Clark, 2004). Since objects do not have to be spatially separated no strategy that depends on assigning a virtual location will work. For example, if two objects are *inside* one another the feature conjunctions cannot be distinguished without considering which object they apply to and where the boundary of the object in question falls.

The problem goes even deeper than just co-locating shapes, orientations, size, colors or textures. The binding problem has to be solved not only for simple 2D shapes, but for three-dimensional objects and, even for *sets of moving objects*. In fact (Alvarez, Arsenio, Horowitz, & Wolfe, 2005; Cohen & Pylyshyn, 2002) found that search for conjunctive features can be carried out over targets during Multiple Object Tracking without any decrement in tracking performance and (Trick, Audet, & Dales, 2003) found that tracking did not impede subitizing (enumerating 4 or fewer moving tracked items). Arguably what is going on in that case is that people select moving objects (using FINSTs) and then, having selected them, they apply their focal attention to those objects to note their color or texture or any other property or combination of properties.<sup>26</sup> The same is true of regions defined in terms of both space and time. The relevant spatiotemporal region is the one traced out by a moving object, not one defined in terms of spatiotemporal coordinates that do not depend on prior detection of an object. Moreover, as noted earlier, perception must provide the information in a form that enables us to conjoin properties across different modalities.

### 3.3.3 Selection and the causal link

The second problem with selection by location arises from the requirement that the link between world and mind be one in which it is primarily the world that initiates the connection – the functions of early vision or sentience are data-driven or, as we say in computer science, interrupt-driven. The possibility that under certain conditions attention might be directed to empty locations (e.g., as it is scanned from place to place under endogenous control) does not help the problem of connecting the world with the mind because *empty places do not have causal powers and therefore cannot have an effect on mental states*. Only *things* at locations can cause a mental event. You can't bind an index to an empty location so you can't evaluate a predicate over it (in the next chapter I will consider whether you can move your eye or your limb to a specific empty location).

---

<sup>26</sup> In MOT subjects generally do not notice such intrinsic properties as color or shape. We assume that this is because such noticings – such property encodings – do require unitary attention which is in short supply during tracking (for reasons that we are still examining). As we saw earlier, tracking objects does not use the contents of the Object Files so is independent of such property encodings.

### 3.3.4 Selection as nonconceptual access

The third problem with selection by location may be the most general and telling of all. If empty places were to be selected – say as consequence of voluntarily scanning attention from one object to another – it would not address the question of how the mind connects with the world. For that the relevant question is: what is selected *nonconceptually*, or without the benefit of a description or encoded properties. What is selected in some specific situation may *in fact* be a location (say some place between two objects through which attention is scanned), just as it may in fact be a speck of dust, the edge of a table, a chair or the capital of France. But because the selection is nonconceptual, even if we happen to select an instance of a P, we do not select it under the category or concept P.<sup>27</sup> So while an empty location might just happen be selected under some circumstance, it cannot be selected *as a place*, since this sort of selection is selection under a concept. At the primitive level under consideration here, what is selected is merely indexed or demonstrated and not conceptualized in any way. The reason for insisting on this stringent constraint is that nonconceptual contact must occur *somewhere* and selection is one of the most promising candidates, for reasons discussed earlier. But when such selection does occur it must not select an x represented as P, or to select x because it satisfies some description P(x), which excludes selection of a place represented as a place because that is selection under a concept, namely the concept of a place. This means that even selection by an indexical locative such as *here*, will not do since that is selection under the concept of location.

Thus the proposal that selecting locations serves as the primitive basis for sensory contact with the world falls short of providing the needed foundation. What is needed is just what FINST indexes provide: a numerically limited mechanism for picking out and preconceptually tracking the enduring identity of things. Because of the causal-connection requirement, the things in question will more often than not be an actual physical object, although it is not encoded as such, which is why I sometimes refer to it as a visual object, sometimes as a proto-object and sometimes simply as a thing (or even as a FING).

### 3.4 Feature placing and sentience

The attempt to understand the primitive connection between world and mind inevitably brings us to the philosophical literature on individuals and on sentience – the sensory contact with the world. Although people who speak of sentience frequently have in mind conscious sensory contact, it is possible to remain neutral on the question of whether one must be conscious of every such contact. To insist on only including conscious access would force us to leave out very many important facts about sensory content that remain below the radar of consciousness.

Austen Clark has recently brought together the literature on individuals – particularly the seminal work of Peter Strawson – and empirical research on perception and attention (as well as some ethologists' writings on animal perceptual orientation). Because Clark's goals –

---

<sup>27</sup> There is an important distinction here that is worth reiterating. Suppose that careful research shows that property P is necessary and sufficient for an index to be assigned. In that case the index carries the information that what is indexed has property P (in the sense of Dretske, 1981). Then one might ask (as Jeff Engelhardt did in a class in which we discussed this topic): Is this then not a case of indexing under the category "*is P*"? Notice that P will almost certainly be a disjunction of properties, quite likely a disjunction of very many properties with nothing in common other than that they attract Indexes (i.e., they need not have a definition independent of the FINST mechanism). But more importantly, there is a difference between selecting all and only things that are *in fact P* and selecting them *as Ps*. In the former case the selection could be made on the basis of any property that is reliably correlated with P (e.g., is nomologically connected with P) so that if P includes "having a local maximum in brightness" the selection is equally correctly described as based on the property of, say, "emitting electromagnetic radiation which does not occupy more than a 5 degree visual angle". On the other hand, selecting things *as Ps* would only select things represented as being P, and if they are not represented they can't be selected in this intentional way. The difference, for those who care, might be described as the difference between *de re* and *de dicto* selection.

understanding the sensory processes that precede conceptualization – are very close to my own, I will summarize some of his views (Clark, 2000, 2004). In the end I will conclude that although the work is well-conceived and insightful, Clark’s acceptance of Strawson’s “feature placing” principles runs him into problems, some of which I have just reviewed, because feature-placing assumes that our initial contact with the world concerns features-at-locations, and therefore that location encoding comes into play very early in the sensory process.

For Clark, sentience is a matter of the sensory experience we have before that experience is encoded in terms of predicates, which he takes to be constitutive of conceptualization, and therefore to understand sentience is to provide an account of what goes on prior to conceptualization. It is clear that what Clark has in mind by sentience is pretty much what I have been referring to as the early vision module – that organ of vision that operates independent of cognition.<sup>28</sup> (Incidentally, Clark’s use of the term “experience”, like mine, does not presuppose that all sensory experience is necessarily conscious).

According to Clark, experience is represented, though not conceptualized. So what exactly is the content of such representations? Clark answers by citing Strawson’s “Feature Placing Language”, a language weak enough to encompass only the representational capacity of sensations and nothing more – so it is devoid of concepts or predicates and cannot represent particulars or token individuals. It is important to Clark that these representations be impoverished. They do not include predicates of any kind, since those are clearly conceptual. Nor does it include identity, multiple or ‘divided’ reference (distinguishing *this* from *that*) or tenses (distinguishing *now* from *then*). Clark summarizes his view as follows (page numbers refer to Clark, 2000):

“The hypothesis that this book offers is that sensation is feature-placing: a pre-linguistic system of mental representation. Mechanisms of spatio-temporal discrimination ... serve to pick out or identify the subject-matter of sensory representation. That subject-matter turns out invariably to be some place-time in or around the body of the sentient organism...” (p 165)

“...there is a sensory level of identification of place-times that is more primitive than the identification of three-dimensional material objects. Below our conceptual scheme – underneath the streets, so to speak – we find evidence of this more primitive system. The sensory identification of place-times is independent of the identification of objects; one can place features even though one lacks the latter conceptual scheme.” (p 144-145)

This feature-placing mechanism serves some of the same functions for Clark that Indexes serve for us; in particular, they allow us to make primitive demonstrative reference.

...Sensory processes can pick out, select, or identify space-time regions of finite extent. If they could not, many of our demonstrative identifications would fail. Seeing which one, hearing which one, and so on all devolve upon capacities of spatio-temporal discrimination. Sensing sameness of location provides the primal form of identification. Language and world come closest to direct contact in the face-to-face encounters of demonstrative identification. This is also the venue where the rational soul and the sensitive soul must meet.” (p 145)

---

<sup>28</sup> In this Clark differs from Peacocke, for whom the preconceptual content of perceptual experience, referred to in the case of spatial content as a scenario, is more fine-grained than usually associated with early vision. Christopher Peacocke (Peacocke, 1992) makes it clear that his “scenario content” is not the same as the representations that are posited, say, by (Marr, 1982). I will come back to this idea in the next chapter. Scenario content appears to be equated with the content of the conscious experience of space.

While there is much in this picture with which I agree (including the emphasis at the end of the last quotation on “demonstrative identification”) I think the basic idea of feature-placing as a way of characterizing nonconceptual selection is mistaken. While there are many arguments against the appeal to feature-placing in the context of sentience (and many have been made by, e.g., Keane, 2004, and the authors of the special issue of *Philosophical Psychology* (2004, Volume 17, Number 4) the arguments against the view that region-selection constitutes the basis for demonstrative identification and for initial selection presented in the previous section are, I believe, decisive. Feature placing requires that early sensory system deliver information that is roughly equivalent to filling out the two arguments in the propositional frame “Property P at location/region R”. Thus the frame requires that a location be specified when any feature is encoded. Clark does not indicate whether either of the two argument positions can remain empty, thus whether a region of empty space can serve in this specification. Empty space seems to me to be ruled out by the requirement that whatever is selected must be able to serve as the causal antecedent of a perceptual event. The arguments given in the previous section seem to me to rule out the feature-placing frame as the basis for preconceptual sensory representation because no selection of a region can be made unless one has made a prior identification of the object in question – it is always the boundary of the object that determines what region is relevant.

Austen Clark’s theory of sentience is important because it lays out some of the boundary conditions that such a theory must meet and it updates earlier work of Quine and Strawson with empirical results on vision and visual attention. All these writers, however, find themselves appealing to *location* as a way of solving the *binding* problem. Quine correctly recognizes that just specifying that several properties are present (i.e., specifying that the conjunction of properties is present) is not enough to bind the properties in the right way. If they are to be conjoined properties we must in addition make sure that they are “superimposed” (Quine’s term) and then bind them by using existential quantification, stating that  $\exists(x)(P(x) \wedge Q(x) \wedge \dots)$ , the way I did in Footnote 5. Quine even says (Quine, 1992, p29) that in doing this “an object has been posited” which sounds like he has embraced our position. But that is not quite so because an object is conceptual, whereas we are still dealing with nonconceptual entities (as we must if we are to have a bridge from the world to conceptual representations). For Quine, as for Strawson, we cannot have an individual without the apparatus of concepts, particularly sortal concepts. This is where one has to bite the bullet and accept that there are things that are individuals, in the sense that they endure and can be continuously identified over certain space-time trajectories, but are not objects in the full physical sense (they could not be in the extension of the concept *physical object*, since that may require appeal to some physical theory within which the concept is defined – and in any case an unlimited number of physical objects cannot even be seen since they are too small or too big). What we pick out and track are not mental representations but things that out turn out to be objects most of the time in our sort of world.

### 3.5 What do FINSTs select? Some consequences of the present view

The claim that we nonconceptually select whatever captures a FINST is a claim that has some rather surprising consequences that should be mentioned. FINST selection is a transparent context – we do not select something as a member of some category or as falling under some concept. When I select, say the rabbit before me, I do not select it under the concept “rabbit” (i.e., as a token of the type “rabbit”) or any other concept. The FINST does not distinguish between selecting a rabbit, a rabbit’s properties (such as its fur, shape, movement, color), a set of undetached rabbit parts, or any other *coextensive* category, because the selection is not made based on a category at all. But if I do not select it under one of those concepts then there is a real

sense in which I don't know what it is that I have selected. I have sometimes referred to the thing selected by a FINST as a FING (a term subsequently adopted by Carey & Xu, 2001). Since I explicitly reject the condition that selection must be conscious (and in that I differ with many philosophers), I needn't even have an experience of what I have selected. Some may find this is a rather surprising position. Indeed, John Campbell has proposed that conscious experience of the referent of a demonstrative is essential for knowing what we are demonstrating. He says, "We need the notion of consciousness in explaining how it is that we have knowledge of the reference of demonstrative terms" (Campbell, 2002, p 136). This may be true for using a demonstrative term in speaking, but in the case of a perceptual demonstrative such as a FINST, we typically do not have conscious experience of its reference unless we apply focal attention to examine it (the issue of the appeal to conscious contents more generally is discussed in Chapter 4). It is only at that point that we may conceptualize it. If the FINST is to serve as the first link in the chain between the world and our concepts then we can't initially "know" the referent. This is the price we have to pay if we are to understand how our conceptualizations can be grounded in causal contact with the world. If we knew what we were going to select, then what we select would fall under a concept. In terms of the previous discussion (Section 1.3.2), the selection would constitute a *test* (or a judgment) that applied the concept, rather than what FINST assignment must be, an *interrupt* in which the selection is causally imposed itself on our perceptual system (hence I frequently speak of FINSTs being captured or grabbed by the perceptual world).

Here is another way to think of the question being raised in this chapter. What does the early vision system – the modular encapsulated system that does not reason from general knowledge of the world, but merely operates mechanically on the inputs from the world – deliver to the mind? Clearly this modular vision system does not deliver beliefs of the form "there is a truck coming from my left" but something much more primitive. Exactly how primitive this message is has been the subject of a great deal of philosophical analysis, including the work of Quine and Strawson cited earlier. It is also the focus of Austen Clark's study of sentience. I suggest that it is even more primitive than what can be expressed in Strawson's "feature-placing language" that Clark adopts (the claim that our initial representation has the form "Feature F at location L"). According to the view I have proposed, what it delivers may be expressed (roughly) in terms of a demonstrative such as "*this*" (although the evidence suggests that there may be 4 or 5 such distinct demonstratives: *this*<sub>1</sub>, *this*<sub>2</sub>, *this*<sub>3</sub>, *this*<sub>4</sub>). In other words it delivers a reference to a selected sensory individual (call it *x*) to which the argument of a predicate can be bound, so that properties may be subsequently predicated of *x* – presumably starting with such predicates as *Object(x)* or *Location(x,L)*.

Of course there must be some empirical constraints on what can in fact be selected in this way. For example what is selected must have the causal power to capture a FINST index. Moreover, there is evidence that not just any well-defined cluster of physical features can be selected and tracked – some, like certain arbitrary parts of things, may not lend themselves to being selected and tracked by FINSTs (e.g., in Scholl, Pylyshyn, & Feldman, 2001, we showed that the endpoints of lines cannot be selected and tracked) and others may be selected but because of the way they move they cannot easily be tracked (for example, vanMarle & Scholl, 2003, showed that objects that appear to liquefy and "pour" from one place to another or that stretch and slink in worm-like fashion can't be tracked). The exact nature of the physical constraints on primitive selection and tracking are empirical questions that we and others are investigating. As scientists we may carry out research to determine the sorts of properties that tend to grab FINSTs. Since these are primitive causal connections we may not be able easily to specify what these connections are connections to in the general case because they could be



connections to any possible link in any causal chain that ends with the appropriate stimulation of the retina (but see below). The category “whatever attracts our attention” may not be a natural kind. Whatever we may discover to be possible properties that cause the assignment of a FINST index, the index itself does not deliver that property or category as part of its function: It just delivers a reference to the primitive selection, to a FING, the way that focal attention might deliver a selection, except that it does it for 4 or 5 individuals. According to the view presented here, it is this selection that enables a reference to the selected FING. Moreover, if the FINST was captured by a property P, what the FINST refers to need not be P, but the bearer of P (the FING that has property P). A FINST refers to something that has properties with the causal power to capture it, even though it need not refer to those particular properties (e.g., it might refer to the object that has a unique brightness without referring to its luminance at all). This is the same as my earlier point that there is a distinction between properties that fix a reference and the properties referred to – a distinction that Kripke relied on in his theory of the reference of proper names..

Notice however, that unlike Kripke’s case of fixation of the reference of proper names, where one can appeal to an initial “dubbing” in fixing the referent, the grabbing of a FINST does not involve an intentional act. Since, according to the present story, establishing a FINST is entirely a causal process, the question arises; Which link in the causal chain determines what is selected? Not the one intended by someone. In vision, for example, the chain includes the light leaving some light source(s), being reflected from some surface(s), passing through the cornea of the eye and stimulating the rods and cones of our retina. Why not say that the light source or some element of texture of the reflecting surface, or the specs of dust in the air through which the light passed, is what the FINST refers to? I said that when properties are encoded, they are encoded as properties of particular FINSTed things – they are represented as  $P(x)$  where P is the property-type and x is the thing referred to by the FINST. So it matters which causal link is associated with what the FINST refers to, since that is what the property P is predicated of. Insofar as selection is a causal process, one might take the position that asking what is causally selected is no different from asking which link in the chain is the cause of the firing of the relevant rods and cones – all links are equally part of the causal story. But that isn’t true of referring. There has to be some unique thing that is referred to. As soon as we have a predicate that specifies a property, some particular unique thing is *represented as* having the property in question. So what determines the particular link in the causal chain that has the predicated property? There are several views about this question which I will not discuss here. It is one of the Big Questions about how reference is naturalized and is beyond the scope of this monograph (and this author).<sup>29</sup>

Whatever a FING is, it clearly does not meet the requirements for individuals as understood by Strawson, Quine, Hirsh, and most other philosophers. This selection does not come with

---

<sup>29</sup> Among the candidates for answering the “which link” question are those that appeal to the conscious content (mentioned earlier), a functional role theory championed by many (it was specifically cited in connection with the very problem of FINST targets by Levine, forthcoming) which says that the referent is determined by the role that the FING plays in psychological processes such as inference and action. Another possibility (recently suggested by Jerry Fodor, Fodor, forthcoming, Chapter 9) is that counterfactuals may rule out all but the correct link in the causal chain. This proposal works because the reference for which we are trying to give an account is a visual reference so only currently visible things are relevant and only links in a causal chain to the FINST from some initial but currently-visible cause have to be considered (which excludes not only the Big Bang and the switching on of the light earlier in the evening; but it allows the light source if it is visible). Such a chain must pass through some property of the referent. Which property or which link in the chain? The answer cannot be determined solely from that one chain – it needs another parameter. According to Fodor’s proposal (which he calls a triangulation), if we consider counterfactual causal chains that end with the same FINST but have a slightly different perspective (a slightly different viewer location) then if the chains intersect they will intersect at the link that is the referent of that FINST.

<sup>30</sup> Luca Bonatti has argued that infants are not sensitive to causality as such, but to continuity in motion and that a better account of the data can be provided based on principles of object-tracking.

conditions of individuation, the mechanism of identity and tenses, and so on. That's because FINGs are not true individuals in the general sense; they are what the visual system gives us in order to bootstrap us into relevant causal contact with the world. This is similar to the situation that faces us in many other logically indeterminate functions such as the visual perception of 3D shapes from 2D images, and many other areas of vision, where early vision appears to reflect the natural constraints of our kind of world so as to accomplish apparently conceptual tasks using wired-in concept-free and inference-free methods (examples of such "natural constraints" were discussed in Section 2.6.1 and will be developed further in Section 5.2.1). There are many examples where the perceptual system seems to be wired (presumably through evolution) so that it represents the information about the world in a way that is constrained so that our representations tend to be veridical most of the time, not in general, but in the kind of world in which we happen to live (our ecological niche). Other examples include such properties as causality and space. Causality, it turns out, is interpreted by the visual system of very young infants who are unlikely to have a working concept of cause – their visual system simply puts into the same category a class of events which we adults would count as exhibiting causality. For example, babies are finely tuned to distinguish between possible causal events such as a moving object hitting another and causing it to be launched into motion, and very similar events in which the timing or the spacing is incompatible with a causal account (Leslie, 1982; Leslie & Keeble, 1987).<sup>30</sup> This, of course, is not the *concept* "cause", which is a difficult concept even for philosophers to analyze, but it is a perceptual category that is close enough to serve in our kind of limited experience outside of science and philosophical inquiry. Let us call these sorts of categorizations proto-concepts because they may serve as the primitive (and possibly innate) nonconceptual *precursors* of later concepts. Other such proto-concepts may include the abstract category *object* and even *animacy* (Scholl & Tremoulet, 2000; Tremoulet & Feldman, 2000).

## Chapter 4. Conscious contents and nonconceptual representation

### 4.1 Nonconceptual representation and perceptual beliefs

In the last chapter I introduced the idea of nonconceptual representation as a form of perceptually-derived representation that does not involve concepts and therefore does not enter into beliefs and thoughts (and probably also memories, since they generally involve inferential reconstructions). Nonconceptual representations have been widely discussed in philosophical circles for a number of reasons. One reason concerns the basic problem that we have encountered in several places earlier in this book: The need for a way to get information from its distal causes through proximal effects (e.g., the retinal image) to perceptual beliefs, the latter being conceptual by definition. The interface is thought to involve a type of information-bearing state whose content is more concrete and detailed than is the content of beliefs, but which nonetheless qualifies as being a form of representation because it carries information about some state of affairs in the world. This form of representation does not represent the visual scene in terms of conceptual categories but is more iconic and uninterpreted. The primary evidence cited in favor of such a form of representation is the disparity between appearances and beliefs (a discrepancy that is the hallmark of perceptual illusions) or between the panoramic, uniformly fine-grained nature of our conscious visual experiences and the relatively abstract, categorical and variable-grain nature of our thoughts, beliefs and recollections. Moreover, the mapping between the proximal stimulus and how the scene appears to us is not fixed, which itself needs to be explained. For example, when we look at an ambiguous figure, such as a Necker cube (e.g., shown in the left panel of Figure 4-1), *something* changes over time as we watch, and that something is not the physical stimulus nor is it generally thought to be our beliefs about what we see. It is what we usually refer to as the *appearance* of the figure or how we consciously experience it. Since the content of our experience seems to be distinct both from the proximal stimulus (the optical projection of the figure on our retina) and from what we believe about the figure, it suggests that we need a vehicle of representation for that type of content.

When we examine a visual scene, the content of our experience is very different from what we know to be the information that enters the visual system. The evidence is very clear that the incoming information is highly incomplete and has a narrow scope (it is literally a moving peephole no more than about two degrees of visual angle) compared with how we experience it (this point will be discussed later in connection with the special case of the experience of space, and is illustrated in Figure 4-3). But the experience also seems intuitively to be different from what we might plausibly capture in terms of the vocabulary of concepts we are likely to have. The argument from the richness of experience compared with the relative poverty of our conceptual resources depends both on how we characterize experience and what we think are the conceptual resources of the mind. But even without considering the fine points of what are reasonable bounds on our conceptual apparatus, it seems clear that we are unlikely to have as many distinct concepts for, say, colors, as there are colors that we can discriminate. Certainly if we consider the number of color terms in known languages we find that the number is actually very small (languages have no more than about 11 monolexic words, see Berlin & Kay, 1969). Yet we can distinguish well over a million different colors (Halsey & Chapanis, 1951), so we are unlikely to encode each of these as a separate concept or code. (Of course there are many more concepts than there are words. But since each word corresponds to some concept, it seems reasonable that each concept is at least a *potential word* in some language. A million different colors is far more than the total human vocabulary so it seems unreasonable to suppose that every discriminable color could have a corresponding concept.)

Notwithstanding such plausible arguments for nonconceptual representations, there are several questions that need to be considered and several tacit assumptions that need to be exposed before the hypothesis that there is a nonconceptual representation of the sort generally accepted in philosophy can be taken as established. The most contentious of these is the assumption that the content of this nonconceptual representation is the same as the content of conscious experience. This view assumes that the content of conscious experience corresponds to a level of representation in an information-processing or functional analysis of the cognitive system. In other words, it assumes that the content of conscious experience is a natural kind for purposes of psychological explanation. I take up this and other issues in the next few sections.

## 4.2 The role of conscious experience in the study of perception and cognition

Cognitive science, and particularly vision science, has had a deeply ambivalent relation with conscious experience. On one hand, the way things appear or what they look like has always been an important, if not the primary source of data, at least for vision science. When one thing looks bigger in one condition than in another or when something looks to be moving faster under one condition than another, or when colors appear different under one lighting condition than another, these are considered primary data to which theories of vision are expected to respond. On the other hand, the content of a person's experience has also proven to be one of the most misleading sources of evidence, because it is not neutral with respect to the theories that the subject holds, whether they be scientific or folk theories. Moreover, what explanations appear most natural is very sensitive to the way we describe our experiences and conversely, the way we describe our experiences (even to ourselves), depends to a large extent on what tacit theories we hold.

The way we describe our perceptual experiences often carries with it the implication that the content of the experience itself *explains* observed phenomena – for example that the occurrence of experience X causes experience Y which then explains some ensuing behavior. There are more or less benign versions of this sort of what might be called intentional causation (also sometimes referred to as psychological determinism, Hochberg, 1968). An important and essentially irreproachable version of this thesis is the appeal to the tight coupling that holds between how a part of a scene is interpreted and how other parts of the scene tend to be interpreted (see Epstein, 1982; Rock, 1997). For example, if you see the edge of a Necker cube marked *x* (in Figure 4-1) as being the top front edge of the figure, then you are likely to see the face of which it is a part as the top face (as in the second panel), but if, instead, you see the edge marked *y* as being the top front edge, then you will see the face that it bounds as the top edge of the figure and the appearance of all the other edges will change so the interpretation of the figure remains coherent. When the percept changes (as it does in ambiguous figures such as in the first panel) the couplings force the interpretation of related parts to change accordingly. This fact has been the basis for a successful technique in computer vision called “constraint propagation” (see the work on the “blocks world” that culminated in the successful system devised by David Waltz, described in, among other places, Pylyshyn, 2003, Chapter 3; Waltz, 1975). (It is also the basis of an approach to models of reasoning using constraint satisfaction, Tsang, 1993).

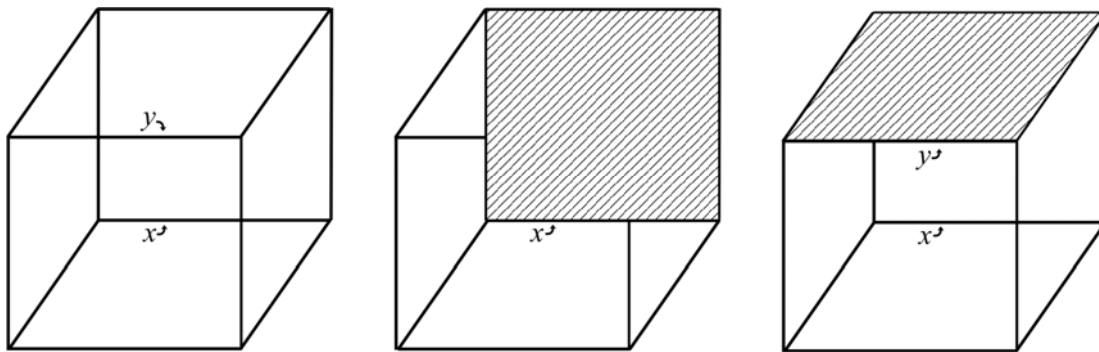


Figure 0-1. The way we see edges is intimately connected the way we see the faces that they bound. These appearances (or “ways of seeing”) form a tightly coupled system. If our percept of one part changes, the appearance of other parts change systematically in order to maintain the coherence of the whole. Panels 2 and 3 show the two versions of the ambiguous figure in the first panel. Notice how the interpretation of an edge is connected to the interpretation of the faces that it bounds and both interpretations change together when the ambiguous percept in the first panel “flips”.

When we speak of labels on the representation of a scene, or indeed when we speak of what a pattern is seen as, we are speaking of the contents of a perceptual representation. In most cases such contents are assumed to be conscious, hence we are speaking of the contents of a perceptual experience. But what exactly *is* the content of a conscious perceptual experience? The content of a belief is relatively clear because beliefs are individuated in part by their contents – that is, we identify a belief by what it is about, or we treat beliefs that are about the same thing as the same belief (I am ignoring for now the fact that beliefs can differ in ways other than in their content – they can also differ in their form). But what about the content of a conscious experience? The situation here is not at all straightforward.

#### 4.2.1 The Contents question

The role that conscious experience can play in vision science depends on our understanding of what such experience reveals. Our first impression is that what conscious experience reveals is both private and obvious to the person who experiences it: if you see something or other, then what the experience reveals is just *what you see*. But if you try to say what that is, you find it is far from straightforward. Indeed there was a long chapter in the history of psychology in the late 19<sup>th</sup> and early 20<sup>th</sup> centuries, where that question was at the forefront of discussions of what psychology is and what it should be (Titchener, 1912; Washburn, 1922).<sup>31</sup> The “Introspective Method” was taught as an objective way to study the contents of conscious experience by turning the mind on itself in a disciplined analytical manner, freeing it to report its own conscious contents, as opposed to reporting properties of the object of our perception (inadvertently being influenced by what you knew about the objects of perception was known as the “stimulus error”). In the end the method failed to provide the foundations for a science of conscious experience, although conscious experience itself continues to be a growing concern in both

<sup>31</sup> For original writings from this era, and earlier, see the interesting web site “Classics in the History of Psychology” maintained by Christopher D. Green at York University, Toronto, Canada: <http://psychclassics.yorku.ca/index.htm>

<sup>32</sup> The Association for the Scientific Study of Consciousness (ASSC) has become a major scholarly society with a large annual meeting where, among other things, the relation between brain and consciousness is discussed and neuropsychologists report various fascinating brain damage syndromes in which patients exhibit disconnects between behavior and conscious contents. See: <http://assc.caltech.edu/index.htm>.

psychology and philosophy.<sup>32</sup> It is not my purpose here to discuss introspection nor to look at the fascinating history of the study of consciousness in psychology. I wish only to point out some of the problems raised by the use of conscious contents as a source of evidence for building theories of perception.

There are two sets of questions about our conscious awareness. One is what might be called the objective scientific question: What are we entitled to conclude about perception from certain perceptual experiences we have? The other, logically prior question, is what the content of our perceptual experience is: What is the thing about which questions of interpretation can be raised? This question is independent of methodological issues concerned with how one should interpret reports of “how something looks”. The question even applies to one’s own (perhaps self-reported) conscious experiences. The question – *what do I experience when I look at this stimulus?* – is fraught with problems. One might reasonably take the position that to ask what we experience is already to ask for a theoretical interpretation, because the experience itself is ineffable; it is part of what (Sellars, 1956) called “the given”. There has been a considerable amount of philosophical discussion of this question. The assumption that one is the infallible arbiter of the content of one’s conscious experience has some serious problems. For our purposes, however, the notion of conscious contents has enough trouble without raising the question of its fallibility, so I will focus on the question of whether one’s conscious experience, or reports of the conscious experience of subjects gathered in the laboratory, forms a reasonable basis on which to build a theory of perception.

Let’s begin by considering what we experience when we look at a scene. Suppose I look at the wall of my room; What is the content of my conscious experience? If I try to describe what I am experiencing I find myself describing the things in the room and their visible properties (such as the color, texture or location). My conscious experience is the experience of the things that I perceive (possibly also what these things remind me of and what feelings they may arouse, but let’s confine ourselves just to the perceptual experience itself). How exactly can I describe what I experience? Does it consist of all the properties that are in the scene, or only those that I notice? Do I experience only what is present in the incoming information, or does my experience also include what the visual system “fills in” and what I infer? For example do I experience the uniformity of the color and lightness of the wall which, as it happens, I know is not in fact uniformly illuminated. Is the uniformity of lightness and color constancy that I am describing an inference or direct experiential content? The lightness you perceive is known to depend on your perception of the location and arrangement of the surfaces in question (Gilchrist, 1977). Also in a typical scene it is rare that I see all of any object because most things will be at least partially occluded by others, even though I do not notice these occlusions unless they are brought to my attention. It thus appears that the content of my experience includes post-constancy and post-filling-in information and therefore my experience relies on more than just the information coming from my eyes.

Let’s continue with this example. There is a picture on the wall. Do I experience it *as a picture* or do I experience what the picture depicts – or both? There is also a calendar on the wall which I see and is therefore part of my experience. Since I forgot to change the page at the end of last month it shows the wrong month. Do I experience it *as a calendar*, and do I experience

it *as* showing the wrong month? Ordinary informal talk is unclear on that issue. If I do experience it as a calendar whosing the wrong month, then other people looking at the wall are unlikely to have the same experience as I do unless they know what I know about the calendar. I look down at my desk and see a sculpture that serves as a paperweight (or perhaps vice versa, depending on how you feel about such sculptures). I see it as a three-dimensional object which has not only a front but also a back and sides and parts that are hidden or occluded by other parts. Do I only experience the front, or only the parts that are not occluded, or do I experience the back and the hidden parts as well? I do not see the back in the sense that I receive no optical information from it, so how can it be part of my experience? Some writers in the Gibson “direct perception” camp claim that we *see* the back of 3D objects as well as the fronts because they are both part of the experience of what Alva Noë (Noë, 2004) calls the “perceptual presence”, which is different from our *knowledge* of the back of a perceived object. This may actually be the more common view. For example, (Block, 1995) says that if you are looking at a set of buildings and then find out that they are mere fronts of a movie set, the content of your perception changes: “The visual experience as of a façade is different from the visual experience as of a building.” If that is the case then conscious content is clearly cognitively penetrable, which affects the role it can play in a perceptual theory.

Many people also say that what we see, our perceptual experience, is “viewpoint independent” which implies that we represent it as a solid without giving a special status to the surface that faces the viewer. By contrast others insist that what we see is just the front of 3D objects and thus what we see clearly depends on our viewpoint. David Marr’s theory of vision explicitly provides a middle ground by proposing what Marr called a “2 ½ D sketch” which is a representation in depth of only the visible surfaces. Which of these is the content of our experience? For Gaetano Kanizsa (Kanizsa & Gerbino, 1982) the perceptual reconstruction of occluded contours in an automatic and cognitively impenetrable process that is part of the process of seeing, and therefore is part of one’s experiential content (I have also defended this view in Pylyshyn, 1999). According to this view, therefore, what we see is not given simply by the incoming information but is a complex output of our early vision system together with some inferences, perhaps from other parts of the scene or perhaps from our knowledge and expectations of what is in the scene. Is all this part of our conscious perceptual content? It is certainly what we mean when we report that something “looks like” so at least in the everyday sense it is part of our conscious content. Where do we draw the line? In (Pylyshyn, 2003, Chapter 1) I give examples to illustrate that the everyday non-technical sense of “what something looks like” is very broad and includes visual puns of the sort popularized by Roger Price in what he called “doodles” (see <http://www.doodles.com>).

In several thoughtful essays, Fred Dretske (e.g., Dretske, 1993, 2006) adds to the perplexity of those who would appeal to the content of conscious experience in building theories of vision, by arguing that we may not always be aware of the content of our experience. That’s because, according to Dretske, there is a difference between being conscious of things and being conscious of facts. That one is conscious of something is itself a fact of which we may or may not be conscious. Dretske gives the example of looking at a wall made of hundreds of orange bricks. Given enough time to scan the wall, does our experience include the experience of *each* of the hundreds of bricks? Dretske claims it does, because if asked whether there was a blue brick among the orange ones we can confidently answer *no*. Dretske claims that this implies that we saw (and experienced) *each* of the bricks since the information that there was no blue brick

depends on having been conscious of the properties of each brick.<sup>33</sup> Yet if asked we might, quite reasonably, claim that we were not conscious of each of the bricks. According to Dretske that just shows that we need not be aware of the conscious content of our perception. Other philosophers have also spoken about the difference between phenomenal and non-phenomenal consciousness (Lormand, 1996), thus further complicating the problem of using conscious contents for theory construction.

There are many examples of our being unaware of information that was readily perceivable and that, by other criteria, was in fact perceived. The question one might ask of each of them is whether they are cases in which we are not conscious of the information, or cases in which we are conscious of the information but were unaware that we were conscious of it. Examples include various cases of apparent functional “blindness”. One of the best known examples is referred to as *change blindness*. In these demonstrations subjects are unable to report the change between two alternating briefly-presented pictures even though the difference between the two pictures is clearly visible when attention is drawn to it (Simons & Levin, 1997; Simons & Rensink, 2005). Another example is *inattention blindness*, in which subjects fail to see a clearly visible feature that occurs at precisely the point where they were visually fixated while they are attending to a more peripheral item (Mack & Rock, 1998). Another such example that is extremely persuasive and puzzling, involves watching a movie with several players who are passing a ball around while the subject is required to count the number of passes. In this example many subjects are unaware of a person dressed in a gorilla suit who walks right through the middle of the scene (Simons & Chabris, 1999). These types of blindnesses appears to involve a failure of information to reach consciousness even though the information is in some sense clearly taken in since it is located in the fovea directly in view. In fact, (Dretske, 2006) claims that the lesson we should take from such examples is precisely that one can be conscious of something and at the same time be unaware that one is conscious of it.

The distinction between consciousness of things and consciousness of facts is similar in spirit, and might perhaps even be subsumed under, the distinction between *phenomenal consciousness* and *access consciousness* advocated by Ned Block (Block, 1995). According to Block, there are two functions of consciousness that should be distinguished – the purely phenomenal function (characterized as “what it is like to be in that state”) and the access function (characterized as states in which information is “poised to be used as a premise in reasoning... the rational control of action... and speech”). These are referred to as P-consciousness and A-consciousness, respectively. These two functions are not only conceptually distinct but they may even involve different neural mechanisms (Block, 2005). According to this view it is possible to have a vivid phenomenally-conscious experience that does not “broadcast” information to other mental processes, and thus that it is possible to have functional access to information that is accompanied by little or no phenomenal experience. Although Block often talks as though these were two forms of consciousness, he does say that he means a “phenomenal aspect” or “phenomenal content” in contrast to “representational aspect” or “representational content” of consciousness. While these two aspects nearly always occur together they are conceptually separable and in some cases empirically separated as when one or the other is damaged by brain lesions. As we have already seen, various types of “blindness” demonstrate information access without phenomenal consciousness. Cases of phenomenal content without access (without

---

<sup>33</sup> In this example, however, judging that there were no blue bricks is likely an *inference* of the form: (1) If there had been one clearly visible blue brick I would have seen it; (2) I did not notice a blue brick, therefore; (3) there was no blue brick. Thus it does not entail that information from each brick was perceptually (consciously) available, only that the perceiver believes that if there had been a blue brick he would have seen it. This is known in the computational inference field as “negation as failure” and is entailed by the “closed world assumption” that is part of the logic programming language Prolog. (see entry in Wikipedia.org).



representational content) are more difficult to find inasmuch as the best evidence for phenomenal content takes the form of verbal reports which *ipso facto* constitutes evidence for informational access.<sup>34</sup> Block himself gives some rather dubious examples of phenomenal consciousness without informational access, which includes orgasms, but in my view the best example comes from split brain patients who are able to carry out tasks without being able to report why they are making the choices they make (discussed in section 4.3.2 below).

Another quite different view of consciousness is provided by David Rosenthal (Rosenthal, 2005) who argues that consciousness consists of being the target of “higher order thoughts” (HOTs) or of having noninferential (unmediated and typically unconscious) thoughts about one’s thoughts. This view, like other views about what it is to be conscious, deals with issues that are beyond the scope of the present discussion. I mention the HOT view, however, because although it is certainly very different from those of Dretske or Block, it does have room to encompass the distinction between conscious experience and awareness of conscious experience or between P-consciousness and A-consciousness. As long as you think that thoughts about thoughts are a real possibility and recognize that they (sometimes) underwrite conscious contents, you might consider cases where thoughts about thoughts do not yield conscious states and also in which conscious states can arise from thoughts about other sorts of mental states besides thoughts (e.g., desires, acts of will, etc). Because all three views allow for a certain degree of independence between qualitative experience and information-processing functions, all these options give one room for the possibility of being conscious of something without being aware of what you are conscious of, or of phenomenal consciousness without access consciousness and vice versa.

My point here is not to advocate a particular way of interpreting the notion of conscious experience, but merely to point out that while the content of experience is important for building theories of perception, it is encumbered with many problems. At the very least the examples above show that whether something is or is not part of the content of our experience is not self-evident, so experiential content is not something we can take at face value on the grounds that since it is *your* experience you alone are the authority on its content. Not only are you not an authority on the content of your conscious experience on any given occasion, but you are frequently demonstrably wrong about what you believe is the content of your experience. Moreover there is no reason why you should be able to say what the theoretically relevant aspect of the experience is and, even worse, you can get it wrong – as we will see in other examples I will provide below.

#### 4.2.2 Conscious experience and public report

Reports by subjects of what something looks like are even more problematic since what people report in an experimental setting is known to be affected by many factors, including what subjects think the experimenter wants (such compliance effects have been called “experimenter demands”), what they believe the task to be (which have been called “task demands”) as well as subjects’ general beliefs and utilities. Every response requires making a decision that may involve weighing the costs (including embarrassment) of different sorts of errors, particularly errors of commission versus errors of omission. For example the work on subliminal perception or “perceptual defense” shows that accuracy in reporting whether one has “seen” a briefly presented word is different for taboo words than for neutral words equated for frequency of

---

<sup>34</sup> David Rosenthal (Rosenthal, 2005, p191-192) argues that it is hard to reconcile cases of P-consciousness without A-consciousness, with the understanding of P-consciousness as “what it would be like” to be in that state, since there is no way to be in that state unless one is aware of it – i.e., without there being some A-conscious aspect we can use to identify those states. This debate is of interest for an understanding of the nature of conscious states, but it does not bear on the current point which is concerned with what we can learn about perception if we set ourselves “conscious content” as an explanandum.

occurrence (Freeman, 1955). These are typically not cases of subjects being disingenuous, but of making rational choices – choices that can be traced to processes described by models of decision-making, such as utility theory and signal-detection theory. Sometimes signal-detection theory can separate contents from reporting biases in a fairly direct way by providing different measures for “response criteria” (the parameter usually written as  $\beta$ ) and for “sensitivity” (the parameter  $d'$ ) in experiments involving thresholds. This is done by taking into account not only the correct responses, but also the relative rate of errors of omission and errors of commission. If the subject has a bias to report seeing something independently of whether there was a signal, then both the hit rate and the rate of errors of commission will increase. Such a bias would be useful if the signal were present on most of the trials or if the utility of detecting the signal was high (e.g., if it signaled danger). If, on the other hand, the subject has a conservative bias, then the hit rate will be lowered but the rate of errors of omission will also increase. This sort of bias would be useful if the signal were present only rarely. These tendencies can be used to separate response bias from the availability of conscious contents (Snodgrass, 2002). Experimental psychology has learned that sincere reports of conscious contents have to be evaluated in relation to other sources of evidence and in the light of developing theories.

Consider, for example, the problem of interpreting such findings as those reported by (Wittrich, 1959). A well-known illusion is that when people walk across the floor of a specially designed room called the Ames room (shown in Figure 4-2) they appeared to change in size.<sup>35</sup> Wittrich confirmed this observation, but he also found that this did not happen when the people were well-known to the observer – e.g., their spouse – even if these people were accompanied by a stranger, whose size did appear to change! Notwithstanding the presumably sincere reports made by the subjects, we may rightly be incredulous of the apparent implication of this finding regarding the malleability of judgments of height, as opposed to the penetrability of reports of conscious experience. The problem is not that subjects are disingenuous, but simply that the lines between what we report and what we believe with great conviction, as well as between what we report to others and what we report to ourselves are not so clear. If, as many have supposed (Block, 1995; Dennett, 1991), part of conscious content is what vision (or imagination) reports to the rest of the mind, then what it reports may be different from the information that it actually possesses. In other words, there may well be a partial dissociation between the content of our conscious experience and the information that is passed on to other stages in mental processing. Sometimes we can show this fairly directly by comparing measures from which the reporting bias has been mathematically factored out, as we do when we use the signal detection measure  $d'$  rather than percent correct. Such measures not only separate what information observers have from what they report to an experimenter, but also from what observers report to themselves – i.e., what they are aware of. More often than not general questions such as this are ultimately adjudicated according to whether a theory that takes certain observations at face value simply misses underlying (causal, functional) principles.

---

<sup>35</sup> It does not reveal a magician's proprietary trick to tell you that the room is actually distorted, having been constructed with one side much lower and shallower than the other. In fact the design specifications are such that rays drawn from a peephole to every visual feature – i.e., corner and vertex (the room has windows) bear the same visual angles to one another as they would have in a regular rectangular room. Thus when viewed through the peephole all the visual cues in the Ames room are identical to those that would have been available in its corresponding phantom rectangular room. Of course it is not possible to build such a distorted room so that the illusion would persist if the viewer was able to move, though this could (almost) be done in an electronic virtual reality display (the reason for the qualifier is that VR displays cannot reproduce all cues exactly; in particular since objects are not actually located at different depths but on the same 2D surface, the eyes do not focus at different depths, resulting in some conflicting cues).



Figure 0-2. From a fixed peephole vantage point the Ames room seems like a normal rectangular room but people look to be different size depending on where they stand. (© The Exploratorium, [www.exploratorium.edu](http://www.exploratorium.edu)).

Focusing on the conscious contents of perception has also encouraged *direct perception theories* (such as those of Gibson, 1979) which claim that perception allows us to directly access (“pick up”) information about properties of the world that are prominent in our experience, such as the property that things have of being suitable for certain purposes – from eating to sitting on (suitability is referred to as having certain “affordances”). James J. Gibson has argued, quite reasonably, that we do not see patterns of light and shadow and patches of color, we see familiar things such as tables and chairs and people. Moreover, we never see just the front surface of objects, we see entire objects and we see them as particular things, such as our car or our spouse, or as having certain affordances, such as being graspable or edible. Although the urge to shun visual representations led Gibson and his followers to embrace what is essentially a behaviorist position, they were right to claim that perception eventuates in the extraction of abstract properties rather than low-level sensory patterns (“sensations”). The moral of this observation should have been that what we see is a reconstruction of the properties of distal objects: we never experience the pre-constancy proximal stimulus. But in direct realism theories (for various modern versions see, Smith, 2003) this is not the moral that is drawn. Rather, these observations are taken to be an indictment of a view that perception begins with properties as described by physical science and constructs a representation of a scene (perhaps in some cases with the aid of inference from general knowledge). Instead they are taken as support for the radical view that the world should be redescribed according to the categories of experience, which are assumed to be the starting point of perception; these are the categories to which perception is inherently attuned and which are “picked up” the way a tuning fork picks up the notes in its immediate environment. To make this picture work Gibson had to deal with the problem of misperception which, in turn, led him the view that theories should be applied to perception in an “ecologically valid” environment (for a critical discussion of these ideas see, Fodor & Pylyshyn, 1981; Pylyshyn, 1984, chapter 6). While it is not usually put in this way, it is the temptation to see the categories of conscious experience as the primitive bases for (or inputs to) perception that has been one of the siren calls of direct realism.

Notice that the position I have been describing in this book bears some similarity to Gibson’s. I too do not believe that we should take the starting point of vision (the nonconceptual first steps) to be sensations, if by sensations we mean consciously experienced colors, shapes, textures, and

so on (or whatever the primitive sensations turn out to be). I have proposed, rather, that the starting point should be nonconceptual, in particular they should be nonconceptual demonstrative references to proto-objects or FINGS.

### 4.3 What subjective experience reveals about psychological processes

In this section I will move quickly through some evidence showing that attempts to infer the nature of psychological processes from the evidence of conscious experience have led us into blind alleys in a number of areas of psychology. I will conclude that the problem arises when one views the content of conscious experience as anything but fallible evidence, which has to be assessed in comparison to evidence from psychophysics and neuroscience.

#### 4.3.1 The illusion of conscious will

The conscious experience of deciding and of willing an action has been called “the mind’s best trick” (Wegner, 2002, 2003). Daniel Wegner has reviewed a great deal of evidence that points to there being large and frequent discrepancies between how and when we have the experience of willing some action and the actual causal antecedents of the action. The research comes from many different phenomena and reveals such things as the following.

1. *The experience of willing an action and the actual decision to act can be dissociated.* It has been shown that the experience of willing an action comes at least 0.3 seconds and maybe even longer *after* the effective decision for the action has occurred, as shown by neurophysiological evidence (these experiments are reviewed in Libet, 2004).
2. *The experience of personal agency or authorship of actions can occur when actions are controlled by someone else.* This is shown in a variety of experiments but can be seen most dramatically in the so-called rubber-hand illusion, in which by the use of mirrors, the experimenter’s (or someone else’s) hand is optically located where the subject feels his or her hand to be, and the manipulation of the fingers in the seen hand is done by the experimenter. In this dramatic illusion, the movements of the hand appears to the observer to be his or her own actions. The illusion persists until some major discrepancy occurs (e.g., the faux hand is withdraw while the subject’s hand remains in place).
3. *The experience of other-controlled action can occur when the action is that of the experiencer.* This is the converse of the rubber hand illusion and has been demonstrated in many controlled experiments, but also occurs frequently in such settings as dowsing (“water witching”), Ouija boards and other “spirit” manifestations where it has been shown that the subject is unwittingly doing the controlling.

A critical aspect of perceived agency comes from the timing of events. Just as we experience causality between objects in such demonstrations as Michotte’s tunnel effect, so we experience ourselves as the cause of some action when the timing is appropriate. Recall that in the Michotte demonstrations, if an object disappears behind an opaque surface (an occluder) and an object appears on the other side at an appropriate time, the experience is that of a single object disappearing and then reappearing on the other side of the occluder; or if an object collides with a stationary object which begins to move, then if the timing is appropriate the experience is that the collision by the first object causes the previously stationary object to begin to move. The same sort of perception of causality appears to work in the case where the first event is the *experience of willing* and the second is some *visible action*; then with the appropriate timing the subjective experience of will is perceived as being the cause of the action.

The conscious experience of will is not exactly the same as the conscious experience that arises in visual or auditory perception – it is not a sensory experience or a sensation – but it is a

conscious experience nonetheless. The person who has the experience reports the clear *perception* that he or she has initiated an action (or in some cases that they did not initiate an action and therefore that someone or something else had done so). These are just the sorts of experiences that make their way into the corpus of data that lend support to one or another theory of perception; they are the “experience *that* ...” something or other has occurred or the “*experience as*” of something or other. So the point here is the same as it was in other cases where the contents of conscious experience are used in building theories of perception; namely, the opportunity for being misled by illusion remains.

#### **4.3.2 Conscious Experience, Interpretation and Confabulation**

Closely related to the illusion of conscious will are cases where observers falsely report the *reasons* for their observed behaviors or the steps they go through in reasoning. The answers to *Why* and *How* questions are often based on one’s conscious experience of psychological processes. The most egregious cases of reports of psychological processes arise in the case of reports of reasoning with aid of mental images, and I will spend some time on this special case later in this chapter. Other cases arise when people are asked to report why they said or did something. Among those investigators who made the most of reports of how and why subjects made certain moves in playing a game like chess or in solving slow and deliberate problems, such as problems in logic, were Allan Newell and Herb Simon, whose work on problem solving appeared in an important book (Newell & Simon, 1972). In those studies they made a great deal of use of “thinking out loud” protocols in which subjects indicated what they were thinking as they attempted to solve a problem, as well as why they considered various options. Even though the problems chosen for analysis were ones that were solved slowly and deliberately and made little use of prior knowledge, Newell and Simon still found that they had to fill in and refine the recorded protocols in various ways. One problem was that subjects rarely disclosed all the moves they considered nor reasons they had for considering and rejecting them. A large number of these intermediate “states of knowledge” went unreported and had to be inferred from other states that were mentioned and from the rational demands of the problem-solving process. Even among the states that were reported many had to be discounted because they played no obvious role in the reasoning path (called a “problem behavior graph”) but seemed rather after-the-fact reconstructions (much the way that recollections are typically reconstructions, as shown in the classical work by Bartlett, 1932). The best such problem-behavior paths were inferred by adding additional sources of evidence, such as eye movements, that proved to be more reliable indicators of what the subject was focused on at various points in time. Thus even under rather the favorable conditions of slow, deliberate and frequently conscious problem solving episodes, the reports of conscious states required a great deal of reconstruction by the theorist.

In social psychology, the idea that we are extremely poor at expressing the processes and the causes of our behavior by introspecting our conscious thoughts is well known (Nisbett & Valins, 1987; Nisbett & Wilson, 1977). Although we think we know why we do things or why we make the choices we do, the evidence shows the contrary (as we already saw in the studies of willing actions in section 4.3.1 above). The reasons we give ourselves and others is more often than not a fabrication based on intuitive folk psychology theories. In addition, the methodology of asking people what they are aware of thinking in the course of planning some action is clearly unsuited for studying such processes as understanding a sentence, where most of the process is not available to conscious scrutiny. In very many cases subjects have no information or conflicting information about their mental processes and when forced to provide reports they simply manufacture (“confabulate”) explanations and rationalizations as best they can.

There has been a great deal of interest in recent years in widespread observations of confabulation, in which people provide descriptions where they did not have the relevant information. Particularly relevant cases are those in which a people do not have access to information about why they made a particular choice (verbally or manually) yet they nonetheless provide a coherent cover story for why they did what they did (see, e.g., Hirstein, 2005). Confabulation is quite frequently reported among patients with dysfunctions that prevent them from accessing the correct information for one reason or another. For example it is often found in patients with large scotomas (blind regions in their vision) that prevents them from receiving information from large parts of the scene. People with these scotomas are often unaware of having blind spots yet (incorrectly) report patterns in the region of the scotoma (in fact everyone has a blind spot where retinal fibers leave the eye, yet most people are unaware of it). There are also some remarkable cases (called Anton's Syndrome) of blind people who insist that they are not blind (denial of obvious impairments is also known as "anosognosia"). These patients guess at what they are shown, and then confabulate elaborate explanations of why they misidentify things by sight and why they bump into things (Hirstein, 2005; McDaniel & McDaniel, 1991). Conversely, there have also been patients who exhibit an even more astounding capacity to make some *correct* judgments of the location and shape of patterns in their ostensibly "blind" field, while insisting that they can see nothing there. These are the famous cases of what is called *blindsight* (Bornstein & Pittman, 1992; Weiskrantz, 1995). Blindsight and other types of agnosias have attracted the interest of students of consciousness because they demonstrate the dissociation of functional vision from conscious visual experience.

An important point in all these examples is that subjects are perfectly sincere in what they report, and do not feel that they are making up answers even though they could not have known the true (and rational) basis for their answers (or their behavior). There are very many things that we do not know but are not aware that we do not know. The contemporary study of what is called *metacognition* is in part about that. Just as we have the *tip-of-the-tongue* phenomenon where we feel we almost have the word we are looking for, so there is the *feeling-of-knowing* (or the feeling-of-not-knowing) which often convinces us that we either know something that we do not or that we do not know something that is just below the conscious horizon (for examples of this sort of phenomenon see the collection of papers in, Metcalfe & Shimamura, 1994). When we think we know something that we do not know, we often engage in confabulation – we make up a plausible story.

#### 4.3.3 Failures of Conscious Access: Split Brains and Split Visual Systems

Confabulation is most clearly illustrated in so-called "split brain" patients – patients in whom nerve fibers (called the corpus callosum) that normally connects the right and left half of their brain are either congenitally missing or was surgically severed to ameliorate severe epileptic symptoms. In these patients, experiences that occur in one hemisphere are not available to processes in the other hemisphere. Since information from the right half of each retina goes to the left hemisphere, control of the right hand is from the left hemisphere, and most language functions are in the left hemisphere, it is possible to set up experiments in which half the brain has the information and the other half has to make a response. Michael Gazzaniga has studied these patients extensively and has reported cases where information is presented to the mute right cerebral hemisphere where it is used to make a response (say a pointing response with the left hand), and then the patients are asked *why* they did what they did. In these cases the left hemisphere that has language must respond – but it does not have access to the relevant information, since it was the right hemisphere that received the information and made the response). In such cases the left (linguistic) hemisphere generally confabulates an answer. Confabulation in split-brain patients has been described extensively (see the summary in

Gazzaniga, 1995; Gazzaniga, 2000). An example that Gazzaniga gives is the case in which different pictures were shown to the two hemispheres of a split-brain patient. The left hemisphere was shown a picture of a chicken claw while the (mute) right hemisphere was flashed a picture of a snow scene. Then an array of pictures was shown to the subject (who is referred to as patient PS) that included a chicken and a shovel and asked to choose one related to the pictures he had seen earlier, making one choice with each hand. The patient chose the shovel with his left hand (controlled by the mute right hemisphere that saw the snow scene) and the chicken with his right hand (controlled by his linguistic left hemisphere that saw the chicken claw), even though the patient could not report seeing the snow scene (since it had been shown to the mute right hemisphere). When asked why he chose the chicken and shovel his left hemisphere replied, "Oh that's simple. The chicken claw goes with the chicken and a shovel is needed to clean out the chicken shed." To account for the way that the speaking hemisphere takes on the task of providing a rationale for the apparent inconsistency, Michael Gazzaniga has proposed a theory that credits the left hemisphere with the task of integrating information, both information that it possesses directly (if it is a left-hemisphere function) or indirectly by observing some of the behavior controlled by the right hemisphere (since, in the case of split-brain patients, it does not have access to what is in the right hemisphere). There were also puzzling examples in which the linguistic left hemisphere was able to give the right answer to something that had been shown to the right hemisphere. Upon careful analysis it turned out that the left hemisphere had observed the response made by the right hemisphere (in one case it had heard a sound that allowed it to figure out that the right hemisphere had touched a brush) and inferred what it had seen, but was unaware that the verbal response was related to this information.

These examples are relevant to the question of what reports of conscious states are about and what they imply with respect to their use in building theories of visual processing. In the split-brain cases, the left hemisphere (which has language and therefore answers the *why* question) sees a conflict between what it experiences and what it infers from watching what the left hand (controlled by the mute right hemisphere) just did, leading to a guess of what might have happened – i.e., a confabulation. Yet patients report that their phenomenal experience in answering the questions in these strange cases is the same as their experience in cases where things are normal (i.e., when they do have the relevant information).

Many of the split-brain examples involve a conflict between two different sources of knowledge (in the two cerebral hemispheres), but there are also many cases where there is no conflict, just the failure of consciousness to access the information that in fact determined an action. This arises frequently under conditions where the motor system is able to act on the basis of information that is unavailable to the conscious recognition system, either due to brain injury or because for some reason the information does not reach the part of the visual system responsible for sensory consciousness. The outstanding example of the former is the famous case studied by Milner and Goodale (Goodale & Milner, 2004), of patient DF who, because of severe bilateral damage in the ventral part of her visual-motor cortex, could not recognize the simplest patterns but could react appropriately and accurately to the same information when executing actions such as adjusting her hand orientation and grasp size while reaching for the article that she was unable to identify or even describe.

This independence of vision for conscious experience and vision for action occurs because the visual-motor system resides largely in the dorsal part of the brain – the part that feeds information from the eye through posterior parietal cortex to the motor system – whereas the conscious recognition system consists primarily of activity in the ventral part of the visual system (which routes information through inferotemporal cortex). In many experiments reported

by David Milner and Melvyn Goodale (Milner & Goodale, 1995) it was shown that in both animals and humans, the part of the nervous system that is in the dorsal visual pathway works differently from the part that is in the ventral pathway. Dorsal processing works rapidly, is more responsive to magnitudes (size, distance, location) and is relatively insensitive to the sorts of visual illusions in which visual context results in an inaccurate experience of size, distance or motion. For example, if a subject reaches to grasp a circle whose apparent size is altered so it appears larger than it really is by virtue of being surrounded by smaller circles (or made to look smaller by being surrounded by large circles), which occurs in the Ebbinghaus or “size contrast” illusion, the grasp-control process is not fooled by the illusion but sets the grasp to the correct size (Aglioti, DeSouza, & Goodale, 1995). In another example, a subject reaches for an object that is displaced during the saccadic eye movement that precedes the arm movement. Because of saccadic suppression, the subject is unaware of seeing any change of the object’s location, yet the (dorsal) reaching system seamlessly corrects for the displacement (Goodale, Pelisson, & Prablanc, 1986).

In studies by (Wong & Mack, 1981), subsequently confirmed by (Bridgeman, 1992) using a different methodology, these researchers showed that the information available to consciousness and the information used by the motor system can be put in direct conflict. The Wong & Mack study involved stroboscopically induced motion. A target and frame both jumped in the same direction, although the target did not jump as far as the frame. Because of induced motion, the target appeared to jump in the opposite direction to the frame. Wong & Mack found that the saccadic eye movements resulting from subjects’ attempts to follow the target were in the actual direction of the target’s motion, even though the perceived motion was in the opposite direction. However, if the response was delayed, the tracking saccade followed the perceived (illusory) direction of movement, showing that the motor-control system could use only immediate visual information, even though the conscious experience is not very different in the two cases. In all these demonstrations it seems clear that the conscious percept differs from the information that the motor system uses in determining actions.

#### **4.4 The phenomenal experience of seeing**

*A note about organization.* In the remainder of this chapter I address the general problem of what to make of our conscious experience of space and of other properties of what mental images represent. Since these topics constitute a central application of the ideas on selection and perceptual demonstratives (FINSTs) I treat them in some detail. However for expository purposes I have divided these topics into two parts. The present chapter emphasizes the role of conscious experience in driving theories in these two areas. The next chapter addresses the same problems from the perspective of spatial representation, discusses some conditions that a theory of spatial representation should meet and offers an alternative account. The account I offer is not a general theory of spatial representation, but an account that deals only with the spatial properties of one sort of representation of space; the representation that we construct when we reason about spatial patterns and relations, which I call Active Spatial Representation or ASPAR. Consequently the discussion of representation of space and other properties of mental images is split between the two chapters.

The conscious experience we have when we imagine something (as when we have a “mental image”) is strikingly like that of seeing something. It is this aspect of the experience that makes it problematic as a source of evidence about the nature of our mental representation. That’s because the experience we have is that of seeing a perceived world and not of our mental state. As with other conscious contents discussed earlier, our visual image is actually the result of many different mental processes, including our perceptual-motor skills, our concurrent



perception of things located around us, as well as inferences we draw from our beliefs about the properties, location and likely behavior of objects we are imagining. Our experience is typically of a stable panoramic layout of spatial locations, some of which are empty while others are filled with objects, surfaces and features that stand in some spatial relation to one another. This is the very phenomenology that leads people to postulate an inner replica of the perceived world and to the belief that this replica constitutes the experiential content of our mental image – a panoramic display that fills the world around us (Fred Attneave called it "cycloramic" since it appears to cover 360 degrees of view, Attneave & Farrar, 1977). If we assume that the content of experience must somehow arise from a representation that encodes that content, and that the representation is constructed from the information we receive through our senses, then there is a problem about how such a representation could possibly come about, given the poverty of the incoming information. The incoming information consists of a small peephole view from the fovea (no more than 2 degrees of visual angle or about the width of your thumb at arm's length) that jumps about several times a second, during which we are essentially blind (the information available to the brain is a familiar story and has been described in detail, see e.g., O'Regan, 1992). So the gap between our visual experience and the available visual information requires some explanation. While there are many ways to try to fill the gap (some of which will be discussed in the next chapter) the natural way, given the form of the experience, is to try to build an internal facsimile of that corresponds to the contents of the experience. In other words we find ourselves postulating a process that takes account of eye movements and constructs an inner picture in synchrony with these eye movements, along the lines shown in Figure 4-3 below.

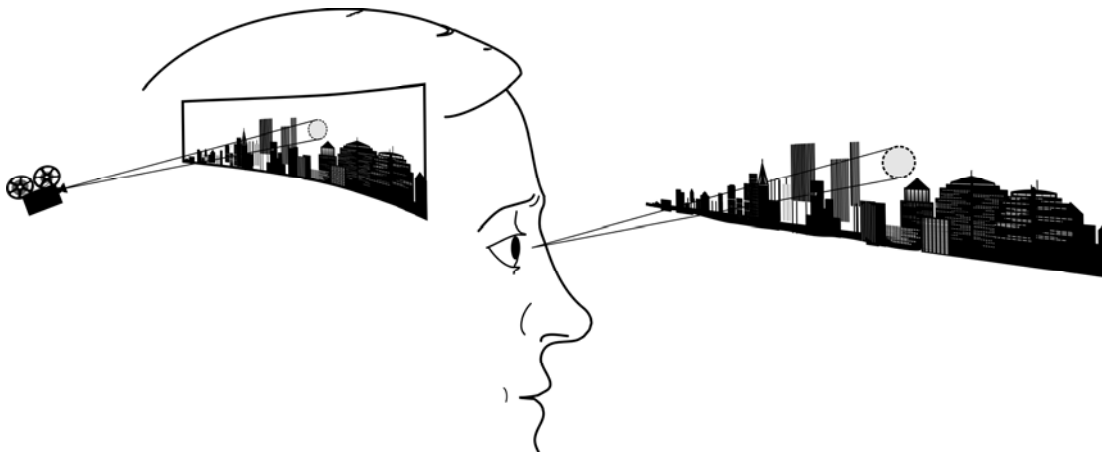


Figure 0-3. The intuitive view of the content of our experience of seeing. According to this view an inner display is constructed by a process that “paints” the display in synchrony with the eye’s scanning of the visual world, thereby achieving a panoramic and filled-in picture of that world, similar to how we experience it (from Pylyshyn, 2003).

But as we now know, this theory is patently false – there is no inner replica or picture of any kind in our head, neither literally nor in any other non-vacuous sense capable of explaining how we represent spatial information in perception and thought. The mistake of reifying the spatial experience in this case is reminiscent of Kepler’s worry (mentioned in Chapter 1) about how we can perceive the world veridically when the retinal image is upside down. Just as Kepler and his contemporaries spent many years looking for a place in the brain where the image was re-inverted, so also have many vision scientists searched for a place in the brain where the fragmentary incoming visual information is completed or filled-in. The experience of visual perception suggests that vision provides a dense manifold of panoramic information, so theorists

have searched for where such a detailed representation might occur in the brain. The answer is: *nowhere*: There *is* no reconstructed detailed representation such as shown in Figure 4-3. The difficulty of accepting this conclusion has gone hand in hand with the difficulty of casting both a theory of vision and a theory of mental imagery in terms other than some form of pictorial or iconic representation, but more on this later.

What has gone wrong that has led so many people to succumb to the “picture” story depicted in Figure 4-3? What has gone wrong is that we are attributing the content of the experience to certain intrinsic properties of a representation (or, more precisely, of the structure or medium in which the scene is represented). But this makes two untenable assumptions. First, it assumes that the content of experience reflects the content of some mental representation that plays a role in the process of perception and imagination. Secondly, it assumes that the content of thoughts or imaginings reveals the structure and properties of the format or medium in the brain in which the mental representations are expressed. I have been discussing the first assumption which gives conscious content a special status over other sources of evidence that I have argued it does not merit. The second assumption is the result of the well-known *intentional fallacy*, the fallacy of attributing properties of what is being represented to the representation itself (as if our representation of a red square were itself red and square). Yet so long as we assume that the form of some mental representation must account for the content of the perceptual experience we are inevitably led to postulate a picture-like representation to match the picture-like experience. Should we, then, discount the experience and start afresh from psychophysical data alone? I will return to this topic below as well as in the next chapter when I consider what a theory of spatial cognition needs to explain – and where I will in fact appeal to some phenomenology to motivate the conditions that need to be met.

#### 4.4.1 Nonconceptual representation in visual perception

As we saw in previous chapters, a theory of perception and cognition needs an ultimate link of some sort with the perceived world in order to ground perception-based mental representations. Furthermore, this link must ultimately be causal on pain of a regress. Or, more precisely, the link must not be conceptual and must not rely solely on the semantic relation of satisfaction. In the previous chapter I offered a proposal for a particular mechanism of selection and reference based on FINST indexing, which we have seen works very much like demonstrative identification. I ended that discussion with the open question of what happens to the rest of the information in a scene, the information about objects and properties that are not indexed. I hinted that we may have to live with the uncomfortable idea that it may be unavailable to the mind, at least at that instant. Such a conclusion seems particularly implausible with respect to visual perception, and in particular with respect to the visual perception of space, because it is here that our phenomenal experience most strongly insists that we have a grasp on space in some sort of bulk manner that is very different from the punctate index-based account I have been offering. This intuition is what people have studied under the heading *nonconceptual representation of space*. The experience of space has been the subject of extensive analysis by philosophers, psychologists and neuroscientists because it offers so much scope for exploring the idea of a different kind of nonconceptual representation – one that departs from the sort of format that seems appropriate for representing propositional attitudes (see, for example, the essays in Eilan, McCarthy, & Brewer, 1993; Gunther, 2003; Luce, D'Zmura, Hoffman, Iverson, & Romney, 1995; Paillard, 1991). The question of how we cognize and represent space will be dealt with in the next chapter. For now my concern is with perception more generally.

As I mentioned above, characterizing the experience of visual perception is a deep and interesting problem on its own. For example the purely phenomenal content of experience may

be relevant to understanding certain distinctions we experience, such as the qualitative difference between vision and mental imagery (Dalla Barba, Rosenthal, & Visetti, 2002) or perhaps between clear perceptions that fail to be convincingly real and vague perceptions that seem very real (a distinction that is orthogonal to perceptual content, as Gestalt Psychologists recognized). Beyond such qualitative observations, it is not clear how cognitive science can build on these ideas, since it is not clear how the detailed phenomenological experience of vision or imagery captures the distinctions and the mental structures required by a causal/functional theory.

Perhaps there is a parallel here with the relation between generative theories of grammar and theories of language learning and parsing. Even while it is clear that the rules of grammar characterize what a speaker implicitly knows, the form of the rules required for characterizing the grammar do not appear to be suited for direct application to parsing or language learning (for more on this issue as it pertains to language see, Pylyshyn, 1973a). In fact it is arguable that the rules (expressed in some generative formalism such as rewrite rules and transformations) are not themselves explicitly represented (Pylyshyn, 1991). Similarly it is not clear that any description of the experience of perception or imagery can be taken as constituting a form of (nonconceptual) representation that is functional in perception.

There are two possible ways of interpreting the claim that even if our perceptual experience were correctly characterized (for example in terms of something like Peacocke's Scenario Content Peacocke, 1992), it need not correspond to the content of some representation that figures in an explanation of how perception works (in an information-processing account of vision).

(1) The first way is that although a scenario may be an accurate description of one's experience, it need not be a good description of any functional mental state. We see many examples of this sort of disconnect between phenomenal experience and functional states in the previous section. But there are even more problems with misinterpretations of experience in discussions of mental imagery. Even when the descriptions of the phenomenology may be correct, their functional significance is problematic at best and often simply irrelevant. To explain, for example, why it takes longer to report details in a "small" image than in a "large" image by saying that the details are "harder to see" in one's "mind's eye", is an example of a spurious explanation based, in this case on the intentional fallacy. This particular way of questioning the role of subjective experience is methodological – there are many sources of evidence that may be brought to bear in the question of how a cognitive process works and subjective experience is just one of them and, like any other source, may have to be discounted in favor of more reliable sources.

(2) A second way of interpreting the claim that even a correct characterization of experience may not provide the right information on which to build a causal theory is more radical. It is the possibility that some of the content of experience may not be relevant to a causal account of perception because a person might *experience* sensory distinctions (differences in qualia) that are not cognitively functional. And that, in turn, suggests that (at least some) differences in qualitative contents may not be functional, or to put it in the strongest possible terms, that qualitative properties may not always supervene on functional or information-processing properties, so there may be differences in qualia that have no corresponding functional differences. A version of this suggestion has been made by Ned Block (Block, 2001).

One of the reasons that philosophy has appealed to conscious experience in characterizing perception is that conscious experience is thought to provide justification for our beliefs. You are justified in believing  $F(x)$  if you can *see* that  $x$  is  $F$  just by looking, where "see" is taken to mean "consciously experience." This may be what is behind John Campbell's insistence (mentioned earlier) that without consciousness we would not know what our demonstrative reference refers

to (see section 3.5.) But there are many ways to justify a belief (even to justify the belief to ourselves) and, given the examples reviewed above, conscious access may not be the most reliable. As Jerry Fodor remarks in a footnote in (Fodor, 2004) that "...the idea that one's justification of a perceptual claim that P, is typically its seeming (its having seemed) to one that P. This doctrine, though venerable, strikes me as confused; in particular, as confusing offering a justification for a perceptual claim with offering a justification for making that claim. Compare: My sincerely believing that P generally justifies my claiming that P; but it's not a reason to believe that P is true (or, anyhow, it's not much of one. Surely it can't be *my* reason for believing that P is true.) Why suppose that the epistemology of perception differs, in this respect, from the epistemology of other sorts of belief fixation?"

Discussion of the content of perceptual experience brings us to the question of the nature of mental imagery which I raise in the next section. In the last chapter I will revisit the question of spatial representation, which many people believe is at the heart of what is special about mental imagery, when I offer a suggestion for how nonconceptual spatial "representation" might arise without any actual internalizing of spatial properties – without an "inner space" of any kind. But for now I will focus on the way that theories of mental imagery are informed (or I should say, misinformed) by phenomenal experience. The point here is that the pull of subjective experience is so powerful and has so thoroughly misled the majority of cognitive scientists (and cognitive neuroscientists) that even patently obvious fallacies go unnoticed.

#### 4.5 The phenomenal experience of mental imagery

If we are tempted by the model of visual perception show in Figure 4-3 above, then we will be equally, if not more, tempted by the view that in the absence of input from the eyes, the inner display in that figure can also be filled from memory or from reasoning (since according to that view there is top-down involvement in painting the inner picture, even in vision). Since, according to that view of visual perception, we have a display surface with the nonconceptual content corresponding to our experience, it would be logical that we might use it to imagine as well as to see. This is indeed the received view in much of cognitive psychology (Kosslyn, 1980; Kosslyn, 1994), neuroscience (see the commentaries appearing with my article in Behavioral and Brain Sciences, Pylyshyn, 2002a) and even a fair amount of philosophy (see, for example, the essays reprinted in, Block, 1981; Tye, 1991). While it is not generally acknowledged (in fact it tends to be vehemently denied), the driving force behind this sort of theorizing is the desire to account for the experience we have when we entertain mental images. The writings on mental imagery begin with the assumption that since the experience of having a mental image is very much like the experience of seeing something, then entertaining an image must also involve perceiving something. And if imagining is seeing something then there must be *something* that one is seeing – there must be something in the head that plays a role analogous to that played by an actual perceived scene (and of course there must be something playing the role of the eye, though that is less often mentioned). Why a picture and not something else? The only other possibility is that there is a replica of a world to be perceived. As Nelson Goodman (Goodman, 1968, p3) said about art (quoting from an unnamed source), "Art is not a copy of the real world ... one of the damned things is enough." This is even truer when applied to mental representations. If it seems unreasonable that there is a replica of a *world* to perceive in the head, perhaps it is possible that there may be at least a *picture* of the world instead.<sup>36</sup> When I am

---

<sup>36</sup> Actually a replica of the world is no more egregious than a picture of the world, given that there is nothing two-dimensional about the experienced image. In fact our images are distinctly three dimensional rather than two, both in their phenomenology and in their psychophysical properties, as I will point out later when I discuss some of the experiments. For some reason it seems less fantastical to ask for just a two-dimensional replica even though all the problems with 3D replicas appear with 2D pictures.

imagining a visual scene it certainly *feels* like I am looking at something and the thing I am looking at looks something in the world. A picture also looks like a world so maybe what we have is a picture. That brings us to the assumption that what we have in our heads (or brains) is something that shares the essential properties of a picture, namely it is a structure that is said to be *depictive* (where the latter is defined in the quotation below).

But it is not enough that we have some structure that looks to be depictive when pictured on paper. The structure itself must be implemented in neural tissue in such a way as to impose constraints like those we find in the world, or at least in a picture, such as requiring that when you scan your eye (or your attention) from place A to place B you must pass through the intermediate (possibly empty) places. Many other properties that we find in mental imagery experiments must also be determined by the structures that underlie the depictive representation. Since most of the constraints that the medium is alleged to impose concern properties of space, I will leave those for the next chapter where I consider the larger question of how one might represent space. What I will not do in the present section is revisit aspects of the imagery debate again. That debate goes back a long way (at least since Berkeley and Locke quarreled over it) and in its modern form it is now nearly 35 old (if we date it from the first salvo in, Pylyshyn, 1973b). While it has changed in emphasis during that time (and has incorporated neural image data) the basic disagreements remain essentially the same. The debate now is fundamentally about whether postulating certain kinds of mental/cortical constructs that are consonant with our conscious experience, constrains the explanatory mechanisms in any way. If it does not, then these assumptions simply appease our intuitions, derived from our conscious experience, without serving any explanatory purpose. This, I claim, is indeed the case for “picture theories” (or “depictive” theories) of mental imagery.

Images are said to be depictive. This is a well-chosen word because it suggests that the relation between mental images and the world is not a semantic one – as understood in linguistics and logic – but one closer to what one might call “resemblance.” Resemblance has a long history in philosophy of mind. It served for Hume as one of the three fundamental principles of association (along with contiguity and causation). But it failed in the end for reasons that are well-known – thoughts can use symbols that do not resemble their referents (e.g., words) and if there is a resemblance the resemblance itself cannot be what determines the reference or meaning (this is not the place to rehearse these ideas, see Fodor, 1965; Fodor, 2003; Pylyshyn, 1984).

One of the people who has tried to be explicit about what it means for a mental image to be depictive is Stephen Kosslyn, as expressed in the following quotation from (Kosslyn, 1994, pp5).

“A depictive representation is a type of picture, which specifies the locations and values of configurations of points in a space. For example, a drawing of a ball on a box would be a depictive representation. ... The space in which the points appear need not be physical, such as on this page, but can be like an array in a computer, which specifies spatial relations purely functionally. That is, the physical locations in the computer of each point in an array are not themselves arranged in an array; it is only by virtue of how this information is “read” and processed that it comes to function as if it were arranged into an array (with some points being close, some far, some falling along a diagonal, and so on). ... In a depictive representation, each part of an object is represented by a pattern of points, and the spatial relation among these patterns in the functional space correspond to the spatial relations among the parts themselves. Depictive representations convey meaning via their resemblance to an object, with parts of the representation corresponding to parts of the object... When a depictive representation is used, not only is the shape of the represented parts immediately available to appropriate processes, but so is the

shape of the empty space ... Moreover, one cannot represent a shape in a depictive representation without also specifying a size and orientation....”

I don't know whether this view is universally received (or even whether it is still Kosslyn's view – see note 49, page 000) but it will serve as a basis for my comments because it has the merit of being explicit. What it defines are the constraints that are assumed to hold by virtue of something's being an image rather than, say, a representation in a compositional system of symbols – i.e., a “language of thought”. Notice right off that what it describes is unabashedly a picture – a 2-dimensional object laid out in space. True, it says that the space need not be physical; it might be only functional. We will see in the next chapter that this idea is a ruse: there is no such thing as a functional space which is capable of explaining the apparent spatial properties of mental images – it is a blank check that can take on any property you wish. It does, however, come close to corresponding with one's phenomenal experience of looking at a picture, which, I suppose, is why we call them “images”. But explanatory adequacy requires that one specify *why* the depictive structure has the properties it has. In particular, there are two very different possible reasons why the representation has the properties it has. One is that this is the nature of the mind/brain – it is part of the relatively fixed architecture of mind or of an encapsulated vision/image module. If the space mentioned in the quotation above were real space such an account would be explanatory: real space has certain properties, including the properties described by the metrical axioms and possibly also the Euclidean axioms that constrain the things that can be represented and the types of transformations or processes that can take place (e.g., moving attention through real space requires that attention “pass over” all the empty places along the way). The second possible reason why the properties described in the quotation hold is that people (i.e., subjects in the experiments) have certain beliefs about what things look like, how they change (e.g., how they move) and how events happen in space and time, and they can use these beliefs to predict or to mimic what would happen in a real situation (e.g., it would take longer to move a greater distance, it would be harder – and so take longer – to see small features than large ones). If the phenomenon holds because a person believes that this is how things would unfold in the world (because of a folk-theory or because of recollecting something similar happening in the past) then the phenomenon does not reveal a property of the mind/brain but only a property of the person's knowledge (often tacit) of how things work in the world. The distinction between a regularity attributed to the nature of the architecture of mind and one that is attributed to tacit knowledge, is about as fundamental a distinction as there is in this field. If the phenomenon is not attributed to some property of the architecture, but to tacit knowledge, then it is in principle changeable by rational means (being told, being shown, etc., any appropriate rationally-connected belief-changing information). The notion of tacit knowledge is one of the fundamental ideas in cognitive science (Fodor, 1968; Pylyshyn, 1981).

#### 4.5.1 Phenomenal experience and explanation: The role of tacit knowledge

One major problem with relying on introspective evidence (even if one is not aware that one is doing so) is that, as in the case of illusions of will and other types of confabulation discussed in section 4.3 above, conscious experience is powerless to tell us *why* something happens. And to the extent that it matters why, we cannot get the requisite answers from our conscious experience. Here are a few examples, intended solely to clarify the difference between an architecturally-based property and one based on tacit knowledge.

Imagine that you are watching Galileo's (apocryphal, as it turns out) experiment atop the leaning tower of Pisa. A large heavy cannonball is released at the same time as a tennis ball. You watch what happens to the two objects as they fall *in your image*. You must press one button when the heavier ball hits the ground and another button when the light ball hits the

ground. What do you think will happen? Well in all likelihood, unless you have studied physics, you would press the button for the heavier ball before that for the light ball.<sup>37</sup> But the critical question is: *Why* did you press the buttons when you did? Does it matter whether it was because of properties of the mind/brain on one hand or properties that you learned in school or believed for other reasons on the other (e.g. watching balls fall in various field games)? Now imagine a person on a bicycle traveling down a hill and then turning around and pedaling back up the hill. Which took longer in your imagining, the downhill portion or the uphill portion? Again, the important question is: Why did those time intervals appear in your imaginings? It should not be hard to think up innumerable such imaginings involving time, each time you are likely to agree that the reason one event takes longer than another is not because of how your mind/brain is constituted, but because of what you know, even if you did not know that you knew (i.e., even if you gave a different answer when asked on a written questionnaire). The use of reaction time in psychological experiments has been a major boon to information-processing theories because it has enabled us to compare the computational complexity (typically interpreted as an indication of the number of operations performed) of processes under different inputs. And yet in this case it seems that it tells us little about the process and its underlying architecture except that it is capable of storing beliefs and drawing inferences from those beliefs and that it is capable of generating time intervals based on independently-computed estimates.

Now try another task using mental imagery. Imagine a beam of blue light and beam of yellow light producing two patches of light on a white surface side by side. Move the patches closer together until they overlap. What color do you see in your image at the overlap? People who have vivid imagery have no problem providing a quick answer.<sup>38</sup> Once again the question of interest is: *why* do they answer as they do? Here's another example. Imagine two identical glasses, one half full of sugar and another nearly full of water. Imagine slowly pouring the water from the water glass into the glass containing the sugar. Does the water in the sugar-glass overflow in your image? The right thing to say in this case is probably, "How should I know?" That answer captures not only one's state of mind, but also acknowledges that what happens in your image depends on what you *know*. In this case the correct answer depends on some sophisticated knowledge about what happens to a solid in solution in a liquid so the correct answer (it does not overflow) would depend on such knowledge or on recollections or on informal folk theory. But whatever answer you give surely depends not on properties of your mind/brain, but on what you believe. Now the reader may well ask how I can be sure that in the above examples the outcome depends on what you believe? The answer is easy: Because you could easily make the outcome different by willing it! It is your image so you could make it do

---

<sup>37</sup> As it happens, this is also what was found in Pisa, not by Galileo but by opponents of the Galilean theory that claimed that acceleration is independent of mass (Kuhn, 1957). Incidentally in this experiment you are also likely to press the buttons after a delay that is a linear function of the distance fallen, which means a constant velocity and not the Newtonian constant acceleration. The dynamics of your mental image have not, it seems, incorporated Galilean physics but remain stuck on the Aristotelian/medieval world view (McCloskey & Kargon, 1988; McCloskey, Washburn, & Felch, 1983). On the other hand, modern sophisticated observers appear to be stuck on Galilean idealizations and have erroneous expectations about motion in air (Oberle, McBeath, Madigan, & Sugar, 2005). Thus what you will imagine in the present example will very much depend on factors that are not related to the architecture of the imagery system (even if there were such a thing).

<sup>38</sup> The chances are good that you gave the wrong answer because few people know the difference between additive and subtractive color mixing. As it stands, the example involves light, which results in additive color mixing, so combining a blue light with yellow light should result in a white patch. But if you were looking through two filters, one blue and the other yellow, you would see a patch of green. For more on this complex but well studied phenomena see (Rossotti, 1983).

whatever you like! If you don't believe me try the above examples making the outcome something quite different: you could make the balls dropped from the leaning tower of Pisa fly away or the fluid miss the container and pour on the floor or the colored light beams mix to form chartreuse or no color at all if you wished!.

I present these examples to illustrate the difference between the two types of causes of imagery processes. While these are not actual published experiments, very similar experiments have been published and discussed in the literature on mental imagery.<sup>39</sup> For example there is a notion of “representational momentum,” hypothesized to account for why in tracking a moving object one generally makes errors in indicating its final position, where it disappears. The idea is that the imagined motion has momentum the way a real moving object does, and like a real object, it does not stop suddenly but continues to move after it disappears. Other such examples are discussed in (Pylyshyn, 2002a).

Having explained the two different sorts of causes involved in imagined processes, we can now look at real examples of imagery that have been discussed in the literature. These generally do not depend on knowledge of physical principles, but on geometrical-optical properties, which seem more likely to be built-in to the visual system. Consider, for example, experiments involving image size (Kosslyn, 1975). In these experiments, subjects are asked to imagine, say, a mouse under two size conditions: (1) imagined next to an elephant so that both are present and fill the entire “mental screen”. In that case the mouse has to be visualized as small – i.e. it has to occupy a small visual angle in the “mental display”, and (2) imagine it by itself in your hand and close up, so it occupies a large visual angle in the mental display. Subjects are asked to report details in the image of the mouse (e.g., does it have whiskers?) under these two conditions. It was found that it takes longer to report such details when the image is small than when it is large. Subjects feel that with the small image they can't “see” the details and have to “zoom in” to see them (and the theory actually postulates a “zoom” operation). So the question is: Is the increased time taken attributable to a property of the architecture or to subjects' tacit knowledge of what it is like to see a small mouse? The pictorialists claim it is the former. The argument given (Kosslyn, Thompson, & Ganis, 2006, p 148) is that “the inhibitory connections in topographically organized areas are typically short, and thus when a lot of spatial variation is packed into a small region strong input is required to overcome the inhibition.” In other words, the visual cortex is limited in its resolving power so you can't get all the information in if the image is small. But a larger mental image is not larger on the cortex. The cortical activity that shows up in PET scans shows at most that a mental image experienced as being larger may be accompanied by activity that is forward of activity accompanying images experienced as small (though even that is not without some question). Thus it may be located in areas where larger retinal images would project from the retina. But even if the locus of cortical activity is shifted when the image size changes, the area of activity for small mental images is not smaller in visual cortex, and so an explanation based on limited cortical resolution is irrelevant. Cortical resolution only applies to the resolution of information originating on the retina, not for information originating internally (from memory) and projected onto the surface of the cortex.

---

<sup>39</sup> Stephen Kosslyn (Kosslyn, 1981) discusses the color mixing example and actually cites empirical data to show that many people give different answers when asked to imagine than when they are asked to provide a verbal response. It's not clear what this is supposed to show, over and above my present claim that people can make their image have whatever properties they wish. I suspect one can get different answers if one asks the question in different settings or in different ways (in a laboratory *vs* a paint studio, inside *vs* outside, in a speeded task *vs* an untimed condition, in a purported IQ test *vs* a children's game, etc): Answers – especially in cases where people do not know the correct answer – can be quickly confabulated (as we saw earlier). They can be answered after giving it some thought, carelessly by free-association or by trying to remember what you might have seen in the past.



For more on the futility of appealing to the neural properties of visual cortex to explain imagery phenomena see Section 4.5.2 below.

But once again I am willing to give the pictorialists all their claims, even though the actual data are problematic in ways discussed in (Pylyshyn, 2002a, 2003), because the problems here are conceptual. What does it mean to make one's image small or large? Can one distinguish between an image being small and imagining something as being small? What do you know about seeing details on a small or a large object? If the size is what makes it easier or harder to see, consider what would happen if we kept the size fixed and manipulated the amount of detail separately. Imagine a medium sized mouse viewed through a pair of steamed or scratched glasses so it is fuzzy. Now imagine the same mouse viewed clearly (you can substitute low-definition and high-definition TV if you like). In which version is it harder to "see" the whiskers? Do we even need to do the experiment? What if we did the experiment and it took more time to see the whiskers on the high-definition image? Would we conclude that the architecture of the visual system has these strange properties? I doubt it. We are more likely to conclude that the subject misunderstood the instructions because what it *means* to be fuzzy is that you can't see the details, and what it means to do the reaction time test is that one is supposed to recreate as closely as possible the phenomena that would occur if one were looking at and seeing the large/small mouse. The details of the neurology of V1 are interesting on their own – e.g., that inhibitory processes may explain the limited resolution of vision – but they do not clarify the problem of the resolution of mental images. That problem arises because one is attempting to match the experience of having a mental image with properties of the architecture of vision when the facts at hand have nothing to do with architecture, but have everything to do with what it means to have a small (or large) image.

The same can be said for the widely-cited study of the "visual angle of the mind's eye" (Kosslyn, 1978). If you ask people to stand close enough to various objects so the objects fill their field of view, you get an estimate of the visual angle of the eye. Now if you ask people to use their imagination and tell you how close they could stand to a car, a horse, a toaster, etc., so that it fills their field of view they also give reliable answers which established that the visual angle subtended by images is similar to that subtended by vision, allegedly showing that the two share a common display. But of course it might also show that subjects know how close they could stand to a car, a horse, a toaster, etc., so that it fills their field of view. The fact that they can't tell you those distances if you ask the direct question is irrelevant here as it is in all the other such cases: One thing that psychologists have learned is that how you ask is critical in determining the answer you get. In this case if you use a different way of measuring the visual angle subtended by images – one that does not invite subjects to imagine that they are getting closer to some object until it overflows their visual field – you get quite different answers. If, for example, you simply provide a task that requires subjects to recall where things are by using their mental image of a room, you find that the visual angle is 360° – what Fred Attneave, whose sympathies have tended towards the picture theory, called a "cycloramic" display (Attneave & Farrar, 1977). The pattern is clear: If you ask subjects to pretend that they are looking at some particular display, or if you present them with a display that they memorize, then you tend to get parallels between seeing and imagining, but generally not otherwise. The tacit knowledge explanation should be treated as the default explanation, barring evidence to the contrary, since that is the way we pre-theoretically understand what imagining something means: by default it is an invitation to put yourself in the position of watching something unfold before your eyes.

Here is another example to which I will return in the last chapter. One of the most widely quoted and replicated results cited in support of the depictive nature of mental images is mental scanning. The finding so impressed the pictorialists that they refer to it as a "window on the

mind” (Denis & Kosslyn, 1999). The typical experiment goes like this. Subjects are asked to memorize a display – usually a map of some fictitious island – until they can reproduce it to within some margin of error. They are then asked to imagine the map and to focus their attention on a particular (named) place on it. Next they are asked to move their attention to another named place. This is done in different ways. In the early experiments subjects were asked to imagine a spot moving from the initial focus to the second named place. In subsequent experiments they were asked to switch their attention or to simply look for the second named place, and in some cases they were asked to report on parts of the map that was off the to side of the imagined region (beyond the visual angle of the mind’s eye). What was found is that the time it took to arrive at the second place was a linear function of the distance between the two places on the map. So the question this raises is Why? The pictorialist has a ready answer – the image is actually laid out in real space in visual cortex so attention (or gaze) travels across it just as it does across real external space and therefore the relation  $time = distance \div velocity$  holds. The account for the case where the item being scanned to is off the image ought to be an embarrassment but it is not viewed that way: there is a story there too (it involves an “image transformation” process – though it’s not clear why that should yield a linear reaction time effect, Kosslyn, Thompson, & Ganis, 2006). Now if you imagine a spot moving across your image of your favorite scene you will notice that it takes more time to go further. But you might also notice that you can make the spot speed up, slow down, back up, hop around, disappear from the scene and do any sort of trick you like. Not only is it your image but it is *your spot* to do with as you choose – the motion of the spot is not constrained in any way by properties of your imagery system. So why do you choose to make it take longer when the distance is greater? Surely it is because that’s what it would do if there was no reason for it to do anything else – *because moving in a straight line at a fixed velocity is what physical things do in real space.*

But wait, you say, the increase in time with distance occurs under many other conditions. It occurs if the subjects are not told to scan, but only to look for or to notice the other named place. But isn’t that the same as asking them to pretend that they are looking at a map? And who would fail to know (or remember) that noticing something further away takes more time – either because it takes time to move your line of sight or your attention? But why do I claim that subjects know that? If you ask subjects what would happen in such an experiment they often say they don’t know. And that’s the right answer – they don’t know what would happen in an experiment. They might even not know what they themselves would do. But that’s just the perennial finding that asking subjects an outright question – especially one about what would happen in an experiment – is the worse way to find out what they believe (witness all the cases covered in Section 4.3 above). You need to see whether different beliefs would yield different results. And you don’t do that by trying to induce strange expectancies by telling unrealistic stories about object movements (such as telling subjects that scanning times would be long for short distances because of some sort of crowding effect or that it would be different for different colored items, Jolicoeur & Kosslyn, 1985).

Another pictorialist defense is to cite an experiment reported in (Finke & Pinker, 1982), in which no scanning instructions were given. Subjects were shown a display of points on a screen. The points disappeared and an arrow appeared. Subjects simply had to extrapolate the arrow to see whether the line would intersect with one of the points that had been there before. Here too, time increased with distance. I leave it to the reader to explain why this does not entail a picture-in-the-head (just put yourself in the position of a subject and ask what you would do and why). They may not have been given instructions to imagine, but they were given a visual memory task which is much the same thing. I think you can probably see where this sort of altercation goes.

Interestingly, pictorialists do not cite the mental scanning experiments we did (actually part of a PhD dissertation by Liam Bannon at the University of Western Ontario) and which I reported in (Pylyshyn, 1981). In one experiment (which was scarcely worth doing since the outcome is pretty obvious if you think about it) we showed subjects a board with a map mounted on it that contained lights and switches. When the appropriate switch was toggled the light that was currently on went off and another went on immediately. Subjects played with this board for a while then were asked to imagine the board and, as in the scanning experiments, to press a button when they *saw*, in their mind's eye, the light come on at the second named place. As you might have guessed, there was no time increase with increasing distance. Why? Because there was none on the situation they were imagining. Notwithstanding such findings, the claim was made (Kosslyn, 1994, p 11) that the research "... showed that imagery is highly constrained. Subjects can control some aspects of processing, such as the speed of scanning, but not others; in particular they could not eliminate the effects of distance on time to shift attention across an imagined object." Yet so far nobody has shown any phenomena that are constrained by the depictive nature of the display and which are not better explained by appeal to tacit knowledge.

It's also interesting that the scanning effect can be made to disappear simply by playing down the importance to the task of moving attention on the image. For example we asked subjects to use their imagined map to do the following. Start by focusing attention on some specified place. Then when a second place is named, they were to say what direction the first place would be from that second place (using a clock face as the way of specifying direction). This is a task which really does require that one focus on the second place in order to use it as the reference point in giving the direction to the first place. But no movement is mentioned. In that case we found no distance effect on reaction time (Pylyshyn, 1981).

But even if you accept the tacit knowledge view, namely that the scanning effect is due to subjects' recreating an imagined state of affairs where attention or gaze is scanned over a scene, there is still one remaining question. How does this simulation create the appropriate observed time delays? Surely it is not the case that a subject simply counts the seconds until the right amount of time has gone by! Why do you get an approximately linear time when (and, I assume, only when) subjects imagine the scanning taking place and not, for example, when they are asked to wait a certain amount of time and then press the button? The answer depends on one's theory of what goes on in the interval. My assumption has been that what goes on is that people imagine the spot or focus of attention being *here* and then *here* and so on until it gets to the target. But does this not require some place for the demonstratives *here* to refer? And does this not require a depictive display? Although as far as I know this question has not been asked – at least not by the pictorialists – and yet it deserves an answer. Precisely this question will be the focus of the next chapter (although the allusion to demonstratives should give the reader a hint that FINSTs are going to play a role in the story).

#### **4.5.2 Does the architecture of visual cortex matter to explanations of imagery?**

I said that the use of tacit knowledge (to simulate what would happen if the event was actually witnessed) is the first line of explanation, but it is not always the last, for there may be other factors involved – some architectural properties often reveal themselves, though not necessarily the architectural properties of the display that are postulated in the pictorialist's canonical story. This sometimes arises in imagery experiments where the task done with imagery is compared with the same task performed with actual vision. Such experiments tend to introduce architectural properties into the picture. I consider one such example because it nicely illustrates three important points: (a) Not all imagery phenomena can be subsumed under the tacit knowledge explanation – many are a mixture of tacit knowledge interacting with some

architectural constraint not obviously related to the task, (b) Here, as elsewhere, the details matter and we have to look at the postulated mechanisms to see what explanatory work they do, and (c) It illustrates how deeply committed some people are to the literal picture theory, so much so that they are willing to ignore some obvious problems in their account in order to salvage the picture-theory of mental imagery. The example I have in mind concerns a low-level psychophysical property known as the *Oblique Effect*.

In the original oblique effect in vision the finding is that it is easier to resolve closely packed lines when the lines are vertical or horizontal than when they are oblique. This general phenomenon has also been found to be true when the stripes are *imagined* at various orientations (or, rather, when memorized bars of different orientations are used in an image recall task such as mentally comparing pairs of such memorized bars for properties like width and spacing, Kosslyn, Sukel, & Bly, 1999). The explanation given for why performance on such imagery tasks is worse when the stripes are oblique is that the imagined stripes are displayed on the visual cortex and it is known that there are more cells with horizontally- and vertically-tuned receptive fields than oblique ones in visual area V1. Now I agree that this does not sound on first hearing like a phenomenon traceable to tacit knowledge being used to mimic perception, but we have already seen examples where the exact wording of the question or of the task made a major difference in the results obtained. The major problem with the tacit knowledge account in this case is that very few people know about the oblique effect even in some informal guise. Of course the fact that people don't explicitly know about the oblique effect does not mean that they cannot *recognize* cases of it – it does not mean that there is no familiarity with how things look when they are oblique, especially since oblique contours are far less common in our world than are vertical or horizontal ones (Hansen & Essock, 2004). We also need to keep in mind that there is a lot we don't know about the oblique effect in general and the (probably many) reasons for it. Notice, however, that the oblique effect is found in the haptic modality as well as in vision (Gentaz & Hatwell, 1998; Kappers & Koenderink, 1999), there are many cases where oblique lines are perceived better than horizontal or vertical ones (e.g., when the spacing is variable and broadband or when the measure is adaptation rather than discrimination, Deford, 2003; Heeley, Buchanan-Smith, Cromwell, & Wright, 1997; McMahan & MacLeod, 2003; Wilson, Loffler, Wilkinson, & Thistlethwaite, 2001), and it seems that the frame of reference for classifying orientation depends on gravity, and so on – all in all not a strong argument for connecting the visual and imaginal oblique effect via properties of a common display in cortical area V1.

But is this result likely due to tacit knowledge? Maybe not – but that is an empirical question, as is the possibility that it is an experimenter demand effect or something entirely different, such as that it's more difficult to imagine oblique lines because they are far less common in the world. What explanation does the pictorialist have for this effect? Here is their most recent explanation (Kosslyn, Thompson, & Ganis, 2006, p 69), "... if the result emerges from the neurophysiology of the visual buffer, it is easily explained... neurons in topographically organized areas are known to have orientation-tuning... and to be less sensitive to distinctions along the diagonal. In addition, at least in area V1 in the cat brain, so-called simple cells (which fire when the animal sees edges and not to complex combinations of features) not only fire more vigorously when horizontal and vertical lines are shown, but also have sharper tuning for horizontal and vertical lines... These results underscore the fact that the oblique effect reflects properties of the neurons that populate the early visual cortex."

So there is the story. There are more finely tuned receptors for vertical and horizontal lines in V1, where mental images are projected as a pattern of activity, and so imagining horizontal and vertical bars get preference in imaginal tasks. But there is a critical assumption in this story which reveals how the seductive picture theory can blind us to assumptions that are essential to

the explanation but remain unstated and unquestioned. The assumption is that a pattern of activity projected onto the surface of V1 from higher cognitive functions, as assumed by the picture-theory is equivalent to the same pattern of activity applied at the retina. But cells that are sensitive to orientation are *sensitive to the orientation of patterns on the retina* – as picked up by photoreceptors – not to patterns imposed on the cells themselves (i.e., to patterns *depicted* on the surface of V1). The orientation sensitivity of cells in V1 is the result of the arrangement of photo-sensitive cells on the retina and how they are connected to the simple cells in visual cortex. The exact form of this arrangement is uncertain – it could be a simple template, as postulated by the organization of simple and complex cells reported by Hubel & Wiesel (Hubel & Wiesel, 1968), or it could be a more complex arrangement such as the wiring of a perceptron-like mechanism – but whatever it is, it is clear that activating an oriented pattern of cells in V1 will not selectively activate orientation-tuned cells. If a pattern of activation, such as a grid of parallel strips, were imposed (“depicted”) on the surface of V1 it would activate all sorts of cells in its path and would not favor different orientation-specific cells depending on the orientation of those parallel stripes. Although parallel stripes on the retina might create parallel stripes of activity on V1 (assuming the retinotopic mapping is accurate), the converse does not hold: activating stripes on the surface of V1 does not produce striped activity on the retina or anywhere that serves as input to orientation-tuned cells.

I have belabored this point because it is a recurring theme in recent picture-theory writings (including the recent overview in, Kosslyn, Thompson, & Ganis, 2006). When certain phenomena of mental imagery, such as the apparent lower resolution of imagery relative to vision reported in (Kosslyn, Sukel, & Bly, 1999), or the longer time it takes to report fine details from a small image, or the “visual angle of the mind’s eye” result reported in (Kosslyn, 1978), the explanation always alludes to properties of cells in V1, in order to support the view that images are projected onto V1. But the property of cells in V1 could not account for such patterns since these properties arise from the way those cells are activated from the retinal photoreceptors; top-down activation from memory or imaging instructions do not produce the same effect as activation from photoreceptors. The only way this sort of explanation would work is if images were projected onto the retina, which so far nobody has had the audacity to propose.

#### **4.5.3 Problems in accounting for phenomenal space by appealing to brain-space**

In addition to the problems raised by attributing properties of imagery to properties of cells in V1, there are even more serious problems with the basic premise behind the evidence cited in favor of the picture-in-V1 view of imagery, namely the assumption that the phenomenal experience of looking at a picture arises from the activation of a pattern on the corresponding display in the brain (specifically on the surface of visual cortex). Many of these reasons are discussed in (Pylyshyn, 2002a, 2003). Here is a quick listing:

##### *(1) Displays in visual cortex (V1) are retinotopic*

Fibers run from the receptors on the retina to cells in the early visual cortex and map spatial patterns in a fairly direct manner (as shown by the activity on monkey cortex when it was made to stare at a flashing pattern – see the photograph of the unfolded visual cortex in Tootell, Silverman, Switkes, & de Valois, 1982) in Figure 4-4. This photograph shows that there is a continuous mapping from activity on the retina to activity on the occipital cortex of the macaque monkey.

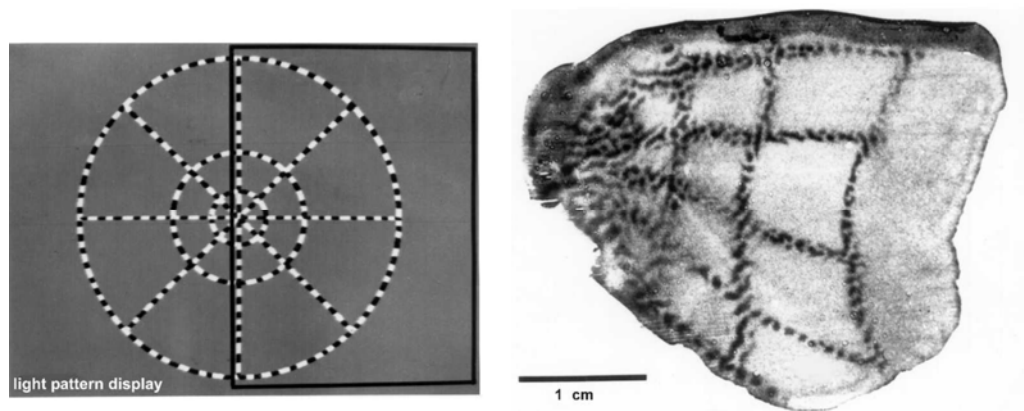


Figure 0-4. Developed “image” on monkey cortex of a pattern of flashing lights shown on the left. The right half of this pattern of flashing lights produced the pattern on the surface of the monkey’s occipital cortex that was made visible in this photograph using chemical tracers (From Tootell, Silverman, Switkes, & de Valois, 1982. Copyright 1982, American Association for the Advancement of Science, used with permission)

Being retinotopic means that, like the retina, the patterns move with eye movements and have a small area of high resolution. If patterns were projected onto the visual cortex in the course of mental imagery and there were eye movements, the interpretation of the patterns would be garbled (and there generally are spontaneous eye movements during imagery Brandt & Stark, 1997). Moreover the mental image is fixed in allocentric space (see Section 5.5.1) – its natural inclination is to remain fixed in extrapersonal space when you turn your head or your body and even when you walk around it! Also as noted earlier, one of the purposes of the depictive image display is to provide a place where the experienced information could be depicted as a panorama with gaps filled, and where mental scanning might take place (since it seems that “mental scanning” can occur even off the central “foveal” region of the mental image, Kosslyn, Ball, & Reiser, 1978).

(2) *Displays in visual cortex (V1) are two-dimensional*

Mental images are three-dimensional, not only in their phenomenology, but also because all objective mental imagery phenomena involving distances or angles produce the same results if they are done in depth as in the plane.<sup>40</sup> To suggest that the display in cortex depicts depth the way we might depict it in a drawing (by using an isometric or perspectival convention) is to miss the basic fact the depictive image is supposed to be the *interpretation* as experienced, not a figure from which such an interpretation is to be derived. If it were not, then we would require yet another form of representation for the interpreted mental image. Since mental scanning and rotation experiments are found in 3D this new form of representation would have to be depictive as well so the 3D problem would still be with us.

(3) *The same cortical display can’t be used by vision and imagery.*

If the same V1 display area were used for vision and imagery then it should be possible to superimpose visual and imagined images to get a composite. While there have been a few claims

<sup>40</sup> Oddly enough, the dimensionality of the display has even been used to support the pictorial view. It has been suggested that the display is able to represent 3D shapes in some unspecified way (orthographic or isometric projection?), but that it is incapable of representing 4D shapes and that this is an inherent constraint of the display itself, which accounts for our inability to imagine 4D objects (Kosslyn, 1981). This ignores the simple fact that we do not know what a 4D object would *look like* which by itself is all you need to explain why we can’t imagine 4D objects. Some physicists think that objects already are 4D (or 6D or higher in string theory) in which case there is no problem imagining them since they look like the ordinary things we see around us!

of this sort, none of them withstand scrutiny (see, Pylyshyn, 2003, section 6.5). If images and percepts both involved patterns on V1 they should be indistinguishable, or at least should be interpretable in similar ways. But images painted on the retina (e.g., afterimages) that are mapped onto V1, work very differently from images created by mental imagery. For example Emmert's Law holds of afterimages but does not hold of mental images. Emmert's law says that the apparent size of a pattern projected onto the retina depends on how far away the background is. For a given retinal image if you look far into the distance the image will appear very large, but if you look at a surface close by it will look small. Although pictorialists are always quick to deny the literal interpretation of a picture displayed in V1, Emmert's Law constitutes a serious challenge to any homeomorphic mapping of retinal onto cortical topography. As long as the image preserves relative size of objects (which has been the explicit assumption of image scanning experiments) then it should be equivalent to a retinal image of a certain retinal size and Emmert's law ought to hold. It's puzzling why pictorialists dismiss this particular problem.

#### (4) *Images are not (re)perceived by early vision*

If images are projected onto V1, then perceiving information from an image should not be very different from reading it from a display. But images cannot be scanned freely, they cannot be freely reinterpreted visually, they do not show signature properties of vision (ambiguity, bistability, visual illusions, apparent motion). Most importantly (and probably the main reason behind all the above problems), is that images are *intentional objects* – they are conceptual interpretations, not raw sensory signals. If an image of a line drawing of an ambiguous figure is imagined, it does not switch between interpretations because it already *is* an interpretation – it is a representation of a 3-dimensional object and it does not change as you look at it (at least it does not change for the same reason that corresponding visual displays change – i.e., because the visual system reinterprets it). If I ask you to imagine two identical parallelograms, one directly above the other, and to connect each vertex of one with the corresponding vertex of the other by a vertical line, no amount of gazing at this in your mind's eye will enable you to see what you would automatically see if you drew it.<sup>41</sup> In addition it seems clear that although you can carry out certain kinds of reinterpretations of geometrical patterns in your image (e.g., you can detect that if you rotate an upper case D by 90 degrees counterclockwise and attach it to a J the result will look somewhat like an umbrella) you cannot do a visual reinterpretation (reperception), as Peter Slezak has shown (Slezak, 1992, 1995); (for further discussion of this issue and other related experiments, see Pylyshyn, 2002a).

There is also evidence (some of it described in Chapter 1 of Pylyshyn, 2003) that images retrieved from memory or created from descriptions, do not function the way displays on the fovea do. Even with visual patterns that are too large to be accommodated in one fixation (e.g., if they exceed the visual angle of the fovea – just under 2 degrees, or about the angle covered by your thumb at arms' length) the part that is off-fovea show signs of already being interpreted. For example if you see a reversing figure such as a Necked cube or an impossible drawing (e.g., the “devil's pitchfork) that is elongated so that some of it is off-fovea, the spontaneous reversal are not observed nor is the conflict caused by the mismatch of two local views that occurs with “impossible figures”. Similarly the famous eye-of-the-needle or anorthoscope presentation, in which a slit is moved back and forth in front of a figure (or a figure is moved back and forth behind a stationary slit) do not yield a true percept with the signature properties of visual perception. Moreover the ability to recognize the shape in the anorthoscope is sensitive to the

---

<sup>41</sup> What this describes is a Necker cube such as the one in the left panel of Figure 4-1. Even if you guessed that it was a Necker cube it would not automatically turn into a 3-dimensional object and involuntarily switch between its two perceptual interpretations (although if you knew about Necker cubes you could switch interpretations voluntarily).

way the figure can be decomposed. If the figure is one that requires a larger number of line labels to be held in memory during the traverse of the slit then it will not be readily seen (see Figure 4-5). In fact quite a few phenomena suggest that off-retinal figures are treated differently – they show more general-memory properties (e.g., they depend on the number of items held in memory) that are the hallmarks of constructed and interpreted figures.

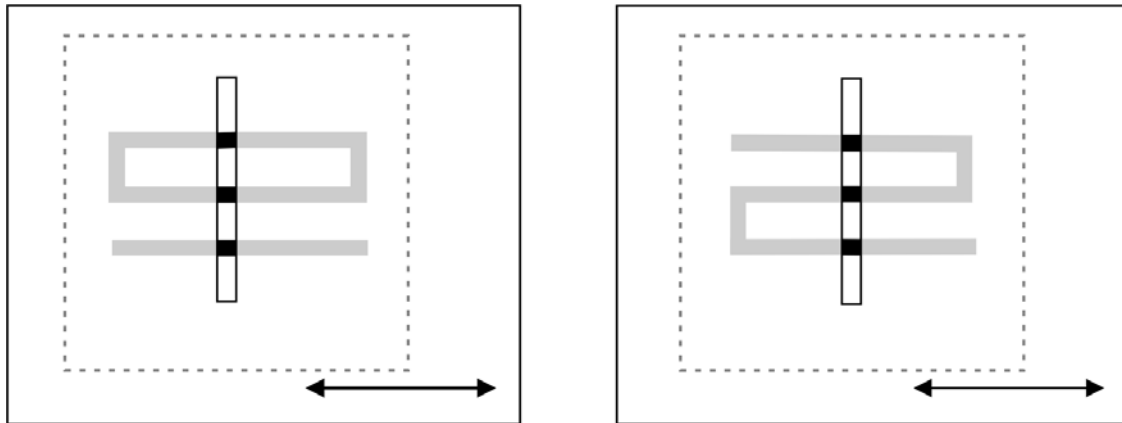


Figure 0-5: An anorthoscope task in which a pattern is viewed through a slit that moves back and forth (or a pattern moving back and forth is viewed through a stationary slit). How many distinct line segments are there? These two displays are identical in terms of their inventory of features so distinguishing which has one line and which has two requires keeping track of which currently-visible segment was connected to which in an earlier part of the viewing sequence. If the figure is rotated 90 degrees the number of line labels that need to be retained is reduced. (From Pylyshyn, 2003, example due to Ian Howard)

#### 4.6 Does phenomenal appearance correspond to a level of representation?

As the examples listed earlier show, what is generally referred to as the “appearance” of the perceived world reflects not only the operation of the perceptual system, but also of our beliefs and expectations and folk theories, and it incorporates these to a much more profound degree than generally believed. Our perceptual experience is the experience of seeing familiar people and things, not of surfaces and textures and shapes and colors. The experience of our own actions reflects what we believe about the agency of the actions, and so on. In other words, experiences are generally of interpreted sensory information.<sup>42</sup> A dramatic illustration of this is what happens under hypnosis where it seems that even the experience of pain (or at least one’s involuntary reactions to it) can be altered. Thus we have every reason to be skeptical about what our subjective experience reveals about the information that is functional in the perceptual process. An even more serious problem with the use of conscious contents for inferring the processes underlying perception is that there is no room in phenomenology-based theories for the growing evidence of vision-without-awareness mentioned in section 4.2.1, including change blindness, inattention blindness, visuomotor control without conscious awareness, blindsight, visual agnosia and disorders of visual-motor coordination and other sources of behavioral and neuroscience data.

<sup>42</sup> There are also intransitive conscious experiences which are not experiences of something or other, but merely floating experiences such as sadness or dizziness or free-floating anxiety. Since I am here concerned with the use of the content of experiences in developing theories of perception I confine myself to transitive experiences (experience as of some sensory input).



I am not suggesting that the perception itself is contaminated by expectations and beliefs – as proposed by many writers in the past, including the “New Look” movement that dominated perceptual theorizing in the 1960s and later (Bruner, 1957), and by linguist Benjamin Lee Whorf (Whorf, 1958) and his latter-day followers (Gumperz & Levinson, 1996) who claimed that language and culture determine how we perceive the world (the Sapir-Whorf Hypothesis). Quite the contrary, I have argued repeatedly that a major part of what we call visual perception is cognitively impenetrable (Pylyshyn, 1999, 2003). The claim I am making now is that the contents of one’s *conscious experience* is the result of these manifold influences, which is a very different claim. Of course if you think that one’s phenomenal experience in seeing the world is constitutive of visual perception, then this does mean that seeing is cognitively penetrable to the extent that the contents of one’s visual experience is penetrable. The widespread assumption that what we see is given by *how things look*, or that how we perceptually experience the world *defines* what we mean by perception, may explain why the recent evidence of unconscious perception or of inconsistencies between how things look and what information is made available to other aspects of cognition (such as to motor control) has attracted so much attention. I am not suggesting that we ignore perceptual experience (as Behaviorists advocated). Not only would this be impossible, given that our conscious experience of seeing, hearing, touching, smelling and so on present deep scientific puzzles themselves, but it is also (and will continue to be) one of the main sources of evidence about perception. The alternative, rather, is to take conscious experience as one of many fallible sources of evidence concerning perception, which may in fact indicate not only what perception yields, but also the process by which we interpret incoming information. This also suggests that we may need to develop methodologies to take cognizance of the multifaceted origins of perceptual experience, the way that signal detection theory provides a methodology for separating sensitivity and response bias in certain decision-making situations (Swets, 1964).

While there are some differences between consciously held beliefs and perceptions and those that are not conscious, these differences appear to be contingent rather than principled in the sense that most of the unconscious ones could have been conscious and in any case they function in similar ways in perceptual information processing. The failure to find a specifically distinct role for the content of conscious experiences, as opposed to the information content of unconscious or unreported experiences in information-processing theories of perception, raises the question of whether the experienced content corresponds to a level or to a type of representation. Representations play a central role in explaining how cognitive processes work, why they lead to the behaviors they do. If there were a type of representation that had a different role, that did not contribute to capturing systematic features of behavior, the question would naturally arise what role it plays in our theories. It may be that conscious contents do not constitute a distinct level of representation because they are a mixture of levels.

Consider, for example, that the inverted retinal image is not part of our experience (at least not *as an inverted image*), although inversion produced by special glasses is. The proximal (retinal) size of a tree, before it is adjusted for the perceived distance away, or the whiteness or color of a surface before it is adjusted for the perceived ambient light (Gilchrist, 1977), and other pre-constancy properties are not part of our experience but they function in information processing the same way as features that we are conscious of (in fact they would continue to function the same way if you did notice them and they became conscious, as happens when one is taught to draw). It is also possible that pre-constancy unconscious states such as those just alluded to are not representations at all, or if they are representations they may be conceptual but subpersonal, in that they involve concepts that do not enter into general reasoning because they remain inside encapsulated modules (see Section 3.2.2). It seems that what we experience is a

mixture of sensory information, subpersonal representations, together with some high-level cognitive recognitions (i.e., familiar people, places, things and events), so it is likely that experiential content consists of various levels of representation.

Although conscious perceptual experience cannot be discounted in the study of perception, neither can one assume that the experience itself is to be taken at face value as an indicator of the nature of a functional mental state – a state that plays a role in the explanation of how perception works. The question of how to interpret a particular observation can only be resolved as we build more successful theories. The situation here is very similar to that which linguists had to face some 60 years ago. Intuitions of grammatical structure, which play a central role in linguistics, similar to the role played by conscious contents in perception, resulted in many disagreements early on. Take Chomsky's famous sentence "Colorless green ideas sleep furiously" which was meant to illustrate the distinction between grammaticality and acceptability judgments. The question of whether this sentence was grammatical led to arguments in which people provided interpretations of the sentence (you can always interpret a sentence, no matter how bizarre its structure). This and other such linguistic intuitions were debated because what constitutes grammaticality, as opposed to acceptability, is not given by intuition alone but must await the development of the theory itself. As generative linguistics became able to capture a wider range of generalizations, it found itself relying just as much on linguistic intuitions. What changed is that the use of the intuitions was now under the control the evolving theories. Even such general questions as whether a particular intuitive judgment was relevant to linguistics became conditioned by the theory itself.<sup>43</sup> So as theories of vision formulate general principles, the theories will direct us to the interpretation of evidence from conscious experience.

---

<sup>43</sup> For example, it was once suggested that grammaticality may not be effectively computable because the judgment of which of the two sentences "I am having trouble choosing among/between P" is grammatical is undecidable (since for a general number theoretic predicate P it is undecidable which numbers satisfy it). One answer to this apparent conundrum was that this just shows that while the choice between "among" and "between" may be taught in grammar classes, it is not a syntactic distinction after all.

## Chapter 5. How we represent space: *Internal vs External Constraints*

*About Chapter 5.* This chapter reviews a wide spectrum of research and makes a particular proposal for how we achieve a sense of space without postulating an inner space. Because the empirical evidence relevant to this thesis covers several strands it might be helpful if I offered a roadmap for this chapter. I begin with a brief review of the problem of representing space and describe two influential theories which propose that general constraints (known as *Natural Constraints* and *Psychophysical Complementarity*), rather than space itself, are internalized. Then I discuss the extremely tempting view that spatial relations are internalized by being mapped onto actual cortical space. I provide a summary of why this view is attractive but ultimately untenable. I then devote some space to a discussion of an option that has frequently been cited, sometimes referred to as “functional space”, which appear to provide some of the benefits of the literal-space model without its problems. I argue that these alternatives either fail to explain the spatial properties, or they reduce to a slightly disguised version of the literal cortical-space proposal. I then list a number of conditions that a theory of spatial representation needs to meet. These are not conditions for spatial perception in general, but only for the representations involved in active episodes of spatial reasoning (what many people characterize as spatial mental imagery, although they apply more generally where there is no conscious experience and where the sense of space is not specifically visual). I refer to this type of case as involving Active SPATial Representations, or ASPARs. At the end of this section I also include a discussion of the much more complicated case of spatial information involved in navigation (often referred to as a “map-like” representation) and suggest that while part of this problem (planning a route) has much in common with the ASPARs, a great deal more goes on in navigation that is beyond the scope of this chapter.

Finally in Section 5.5 and 5.6 I get to the proposal for how ASPARs get their spatial properties (i.e., the sample of 6 conditions of spatial representation laid out in section 5.4.1). It is an externalist proposal which claims that the spatial properties represented by ASPARs are inherited from the spatial properties of the concurrently perceived world (I call this idea the Projection Hypothesis). The objects of thought in ASPARs are associated with particular things in the concurrently perceived world by the mechanism of FINST indexes, and by this means the spatial layout of these things becomes available to perception. Thus such properties as the apparent analogue representation of distances and the configurational properties of sets of objects can be accessed perceptually since they are now out in the world rather than in the head. An example of the way this hypothesis works in vision is provided. The next step is to suggest how the projection hypothesis works in other modalities – how one can index and bind things in nonvisual modalities. To arrive at that point I make several detours to introduce a number of findings from perception and neuropsychology. These include a number of experiments showing that concurrent stimulus input plays a role in spatial reasoning and that because of temporal lags this influence endures slightly beyond the duration of the stimulus itself, that deficits such as unilateral spatial neglect can be modified through concurrent stimulus inputs, that the ubiquitous presence of mechanisms of coordinate transformation in the brain makes it unlikely that we represent spatial properties in a single allocentric frame of reference while at the same time it makes it possible for proprioception and motor control to operate as though it could access places in an allocentric frame of reference, including places of intended (but not executed) actions. In the end I suggest that while this proposal is somewhat speculative, the evidence I have marshaled is compatible

with the thesis that the spatial character of representations of space derives from the concurrent perception, through many modalities, of objects laid out in space around us to which we can bind the objects of thought and can then inspect them perceptually.

## 5.1 What does it mean to represent space?

Any analysis of the mind-world connection will sooner or later have to face the problem of how a mind is able to cognize space. Indeed, the problem of “spatial representation” is one of the deepest problems in cognitive science (Eilan, McCarthy, & Brewer, 1993; Paillard, 1991). Arguably the problem of understanding how mental imagery functions is to a large extent the problem of how we represent and cognize space. One of the difficulties in understanding our commerce with space is the fact that it is so extremely intuitive to us that it is unclear what we mean when we ask how we cognize space, how we deal with spatial properties that we learn from our sense perceptions. It seems obvious that space is that unique three-dimensional receptacle in which objects reside, and that spatial relations are there for us to see and experience without the need for any inferences.

In addition, our conscious experience of space is all-pervasive and fine-grained; we experience ourselves as being totally immersed in the space around us which remains fixed as we move through it or as objects other than ourselves move through it. Our spatial abilities are remarkable and have resisted a causal explanation despite the efforts of some of the best minds over the past two centuries. We can orient ourselves in space rapidly and effortlessly and can perceive spatial layouts based on extremely partial and ambiguous cues. We can judge depth with extremely subtle cues, either through retinal disparity or through parallax based on extremely small movements. We can recall spatial relations and recreate spatial properties in our imagination. We can move through space rapidly while avoiding obstacles. Animals and insects low on the phylogenetic scale, who may not have concepts, inasmuch as they arguably do not have the power to reason about things that are absent from their immediate perception, exhibit amazing powers of navigation which proves that they have *quantitative* representations of the space through which they travel and that they update these representations continually as they move through the space or interact with it in various sensory-motor modes. Although perception science is arguably the most developed of the cognitive sciences there are many areas where it is far from clear that we have even posed the problems correctly, and the problem of spatial cognition strikes me as an extremely likely candidate for one of those problems.

To our modern sensibilities it also seems that space consists of a dense array of points which can be connected by straight lines. But these notions, which have been enshrined in our view of space at least since Euclid, may not be the right notions in terms of which we ought to describe how we perceive space and especially how we represent space in our mind when we think about it or imagine events taking place in it. But what does it mean to say that this is not the way we cognize space? How can it be that the way we register space does not give a privileged place to points and lines?

I begin by trying to outline the nature of the problem that faces us. What does it mean to see the world as being laid out in space? What must the architecture of a mind be like that can do this? Given the patterns of energy that impinge on our sense organs, what must the mind do to create the particular understanding of space that we have, and the special skills that we have for reasoning about it, perceiving it, and moving through it? Many of our greatest thinkers have sought to answer such questions, from the early Greek thinkers like Euclid through the Enlightenment, with names like Kepler and Descartes, and later Poincaré and Kant. The problem fascinated thinkers like Johannes Kepler, who (as we saw in chapter 1) was one of the first to recognize what we now take for granted, namely that: (a) the retinal image plays an important

role in the causal chain by which we come to know about space and that (b) the gap between the retinal image and the apprehension of space would not succumb to the same style of geometrical analysis that worked so well in explaining the connection between the light, the objects, and the image on the retina (Lindberg, 1976). René Descartes' arithmetization of geometry was one of the seminal accomplishments in understanding that the problem had a formal structure that was amenable to rigorous study. Then in the 20<sup>th</sup> century several great French natural philosophers were struck by the problem. Henri Poincaré (Poincaré, 1963/1913) was one of the most important of these and I will return to his views below. The problem of developing a sensory-based Euclidean geometry was raised again by Jean Nicod who, in the 1930s wrote a dissertation entitled "Geometry and the Sensory World" which laid the groundwork for a very different way of looking at this question (Nicod, 1970) and which had an effect on the thinking that led to the FINST theory.

For Nicod the problem was that the basic building blocks of the Euclidean (and Cartesian) view are *points* and *lines*, and a way of constructing figures from them, together with the relation of *congruity*, none of which seemed to Nicod like the sorts of things that perceptual systems are equipped to detect and to use – they are complex *types* that collapse collections of sensory experiences into categories that make the statement of geometrical principles simple at the cost of making their connection with sensory data opaque.<sup>44</sup> Nicod pointed out that since there are very many models of the Euclidean axioms (the Cartesian mapping of points in space onto triples of real numbers being the best known) we should seek the one that captures Euclidean spatial properties in terms of primitives more suited for creatures with sensory systems like ours. After considering a variety of such possible primitives, he developed several "sensible geometries" based on the geometry of volumes and of volume-inclusion (or what he called "spatio-temporal interiority") and argued that this basis is closer to our sensory capacities than one based on points and lines (one reason being that volume inclusion is invariant with viewpoint so it can be sensed as we move through space). With the addition of a few other novel ideas (such as the idea of "succession" and of "global resemblance") Nicod set out a new direction for understanding what space might consist in for a sentient organism. While in the end he did not succeed in developing a complete formalization of geometry based on these sensory primitives, he did point the way to the possibility of understanding sense-based space radically different from the Euclidean, Cartesian, and Kantian approaches that seem so natural to us. If Nicod had been able to carry out his program (he died at the age of 33 shortly after writing the thesis on sensory geometry) he might have provided a set of tools for viewing space that would have been more useful to us than the view that is thoroughly embedded in our way of thinking. But he did show us that thinking in terms of points and lines may not be the only way and indeed it may not be the most perspicuous way for cognitive science to proceed in studying the psychological nature of space. This theme is one with which the present book has some sympathy, insofar as it begins not with points in space nor with lines and shapes, but with the notion of an object (or proto-object).

## 5.2 Internalizing general spatial constraints

Since Watson's identification of thought with subvocal speech there has been a strong mistrust of accounts of mental phenomena that appeal to direct internalizations of external

---

<sup>44</sup> In talking about Nicod's views, Bertrand Russell put it this way: "... the formation and growth of physics are dominated completely by the pursuit of simple laws or, better, by the simple expression of laws. This expression can in fact only be obtained by marking complex things by simple names. For nature is constituted in such a way that it is not the simple things that enjoy simple laws, and so, in order to simplify the laws, we must complicate the meaning of terms." Jean Nicod attempted to find simple things in geometry at the cost of complicated the statement of the geometrical principles. (Nicod, 1970, Introduction by Bertrand Russell)

properties.<sup>45</sup> I share this mistrust and continue to believe that cognitive theories that exhibit this sort of interiorizing of externals betray our latent behaviorist instincts, our tendency to focus on observables even if they are hidden just below the skin. But not all internalizations are misleading – in fact being intentional organisms entails that *in some sense* we have internalized (i.e., represented) aspects of the world and thus that our behavior is not continuously under the control of stimulus features, as assumed by Skinner (for a devastating critique of this entire behaviorist project, see Chomsky, 1957). In earlier chapters I argued that in addition to representations that are related to what they represent by the semantic relation of *satisfaction*, we need a more direct or causally-based relation. I suggested that other sorts of internalizations besides conceptual ones play a role and therefore that we could learn by taking a second look at the general issue of internalization. In what follows I very briefly sketch two approaches to the question of internal constraints on spatial representation that, unlike the internal display or picture-theory, provide useful ways of looking at aspects of the problem.

### 5.2.1 Marr's principle of Natural Constraints

Computational vision, perhaps more than any other approach to understanding visual perception, has faced squarely the problem of the poverty of stimulus information in relation to the richness of our perceptions. As is well known, the mapping from a distal scene to the retinal image(s) is not invertible – an unlimited number of distal patterns are mapped onto the same proximal pattern, so the proximal pattern does not univocally determine the distal pattern. Yet the visual system computes a univocal inversion – we almost always see a unique spatial layout despite the ambiguity of the incoming information. How we can do this has been the subject of speculation for many years, with the prevailing view in the second half of the 20<sup>th</sup> century being that visual interpretation depends on our knowledge of how things are likely to be in the world, and in particular knowledge and expectations of the particular scene in question. James Gibson questioned this assumption, insisting that the information was all there in the ambient light if we only looked for it in the right way. But it was David Marr (and others working on human and machine vision, Horn, 1986; Koenderink, 1990; Rosenfeld, Hummel, & Zucker, 1976) who made the case convincingly that vision does not need (and indeed, is *unable* to use) information from our general store of knowledge in interpreting a scene (the argument is also made at some length in Pylyshyn, 1999, where the reader can also find counterarguments and discussion). Rather, the reason that vision comes to a univocal interpretation of spatial layouts is that it is unable to entertain the many alternative hypotheses compatible with the sensory evidence. And the reason for that is the existence of what Marr called “Natural Constraints” which consist in very general constraints on interpretation and representation that were compiled into the visual system through evolution and that reflect the nature of the sort of world we inhabit. It's not that the visual system knows that the scene before it consists of rigid objects, but rather that it is so constituted that (to a first approximation) only interpretations consistent with the rigidity of most objects are available to it. If you knew that the objects were not rigid it would make no difference to the interpretation that vision would provide.

This idea, though not entirely unprecedented, was revolutionary when combined with a program of research in computational vision. The task then became to uncover the various natural constraints that are built into vision and to show how a system that respected these constraints could see spatial layouts the way humans do. This led to a series of projects typically

---

<sup>45</sup>Those acquainted with epistemology will find this use of the term “internalize” and “externalize” somewhat deviant, for I am not concerned with epistemological questions of justification (Goldman, 1986) where the terms frequently occur. Nonetheless the terms seem to be appropriately descriptive for present purposes.

entitled “structure from X” where the X’s are such sources of information as motion, shading, stereo, contour shape, and so on. This is a sense of internalizing of “constraints” that is both theoretically plausible and empirically validated – at least in many cases. The approach is closely related to a similar goal in linguistics where both language learning and sentence comprehension are underdetermined process: The data on the basis of which languages are learned and on the basis of which sentences are parsed are similarly impoverished. Indeed there is a mathematical proof that you could not univocally infer a grammar (any grammar) based on samples of sentences alone (Gold, 1967), and there are also proofs that some particular principles of, say, English arise because without those principles certain aspects of the grammar of the language could not be learned (Wexler & Cullicover, 1980). What is assumed to enable the learning of a native language in the face of the indeterminacy grammatical rules are the innate brain structures described by Universal Grammar which prevent the infinite number of humanly inaccessible languages from being learned or the similar infinite range of sentence parsings from being considered. Similarly the interpretation of visual signals is constrained by internalized natural constraints.<sup>46</sup>

The question of whether these constraints allow for the representation of spatial information without conceptualization is an interesting one. If the representation of space as we experience it were achieved without first carrying out inferences (which require conceptualization), it would be a good candidate for a nonconceptual form of representation. This is very close to what the natural constraints idea proposes. It claims that the representation of space is achieved without inferences and therefore without necessarily conceptualizing the sensory information. A major question that this story raises is whether such natural constraints apply to representations constructed in thought (as mental images) as well as in perception. Is the fact that we cannot imagine a 4-dimensional space related to the fact that we cannot perceive one? Perhaps. Or perhaps it is simply because we don’t know what such a thing would look like. In general it is not clear how our ability to imagine spatial layouts can be explained by general constraints of the sort that are postulated in vision. While thinking is, by definition, carried out with conceptualized representations, yet there is no principled reason why a nonconceptual representation could not play a role if the cognitive architecture made this possible. We can, after all, make use of external representations such as diagrams and drawings, so why might such things not be constructed in thought? As we saw earlier, one of the problems with such a proposal is that imagination is creative in ways whose boundaries are unknown. While we may not be able to imagine a 4 dimensional space, we can readily imagine objects moving in very nearly any way at all, with or without maintaining rigidity, with or without obeying the laws of physics or the axioms of geometry. And this plasticity of thought is a major problem for any internalization theory, as we will see.

### 5.2.2 Shepard’s psychophysical complementarity

Roger Shepard (Shepard, 2001) has take the idea of internalized properties and principles even further. Citing the example of the circadian rhythm which internalizes the earth’s daily light-dark cycle, he argues that many reliable generalizations of human behavior can be traced to our internalizing universal properties of the world. His general argument is based on the

---

<sup>46</sup> It is worth pointing out that even though the existence of such constraints in both perception and language is well accepted, there is some disagreement as to whether these reflect properties of our sort of world or whether they should be attributed to some sort of innate optimization (or “regularization”) process in the brain (see, e.g., Poggio, Torre, & Koch, 1990). For our purpose it does not matter – in either case there are constraints that prevent all logically possible analyses from being actual candidates and it is the existence of these constraints that allows otherwise indeterminate problems to be resolved uniquely (in vision it is the analysis of a 2-dimensional signal as originating from a 3-dimensional world and in language it is the acquisition of a grammar from an impoverished set of exemplars).

evolutionary advantage of being constructed in that way – a sort of Leibnizian “pre-established harmony” between world and mind. But such an argument should lead us to expect that universal physical properties would be internalized, for what is more important than correctly anticipating where and when a rock will fall on you? Yet this is not the case; physical principles do not seem to have been internalized the way geometry has in either in vision or in thought, and especially not in infant cognition despite infants’ ability to recover many abstract categories such as causality or agency (Leslie, 1988). What appears to be internalized, according to Shepard, are “principles of kinematic geometry” or of transformations of objects through space. Because of these internalized principles we tend to perceive objects as traveling through the simplest geodesics in a 6 dimensional space (3 dimensions of translation and 3 of rotation). Shepard presents some interesting examples of this principle, involving apparent motion of asymmetrical 2 dimensional forms that tend to be seen as traveling in 3D according to a screw transformation, as predicted by the “simple geodesic” story.

These are interesting and suggestive ideas and if there is anything to the internalization principle these are certainly good candidates. But neither Marr’s Natural Constraints nor Shepard’s geodesics constitute internalizations of space in a strong sense that has been assumed by many psychologists, but rather they are principles of abduction that determine which hypotheses the organism is able to entertain. In that respect these constraints are like Universal Grammar. The alternative to such constraints is the conceptual story: rather than internalizing properties, we learn about them (by induction) and then we draw inferences from them. In the end this cannot be the case for all cases of spatial properties, for reasons that I have discussed, namely that our beliefs must eventually make contact with the world, so the initial step must be causal and therefore the only open question is how this happens. Internalizing is a way of incorporating *principles* or *constraints* in a nonconceptual way. It is not surprising, therefore, that the only candidates for this sort of internalization are constraints that apply to modular parts of the perceptual systems where general reasoning does not occur because processes there are encapsulated from the rest of cognition.

While the Shepard type of constraint does not entail a particular format or medium of spatial representation, some people have taken that proposal as evidence that the properties of Euclidean space are somehow internalized as an inner space (for example, a space defined by states of assemblies of neurons – as proposed by some of the commentators on Shepard’s paper, such as Edelman, 2001). If they are, then there is little evidence of their operating in thought – particularly in mental imagery. We can easily imagine objects breaking kinematic principles, traveling along nongeodesic paths and violating just about any principle or constraint you can think of.<sup>47</sup> And this, as I remarked earlier, is what eventually leads most inquiries into the cognitive architecture to the view that the mind is much more like a Turing Machine, than like any multidimensional space or connectionist network (Fodor & Pylyshyn, 1988).

### 5.3 Internalizing spatial properties by mapping them onto an inner space

Before proceeding with this survey of approaches to representation of space I need to mention two proposals that arise primarily in the context of the mental imagery debate. They

---

<sup>47</sup> As noted in the previous chapter, a major factor controlling how we imagine things is the implicit requirement of the imagery task, namely that it is about recreating how things would look if they were seen. Thus one apparent constraint on mental images is that they are viewed from a single point in space, as opposed to several places at once (like a Picasso painting) or from no perspective at all. This is surely because we do not know what it would *look like* if viewed in any other way. To imagine that the world is some other way (e.g., four-dimensional) may in fact be possible if we put aside the requirement that we imagine seeing it that way. For example it is not inconceivable that one could learn to solve problems in 4-dimensional space (perhaps with some virtual-reality training). Some people claim to be able to *reason about* such objects, if not to visualize them (Hinton, 1906).



both involve strong forms of internalization in that they propose that we cognize perceived or imagined space by mapping it onto an internal space. The strongest form of this proposal actually locates the space of mental representation in the literal space of the surface of the unfolded visual cortex. The weaker form, that most people cite when they are pushed on the question of whether there is a literal space-in-the-head, is the idea of a “functional space”. The former proposal (brain space) runs into both conceptual and empirical problems and the latter proposal (functional space) is either a confusion about the explanatory value of the appeal to a “functional space” or it is a proposal that spatial properties of represented space arise from the nature of the architecture of the mind/brain, which is a slightly disguised form of the brain space proposal. Because these proposals illustrate several points about the explanatory role of representations I will take a few pages to discuss them before proceeding to the present externalist proposal.

### 5.3.1 Brain Space

If the perceived or imagined space were mapped onto a literal space – a neural layout in the brain – then we could have an account that explained many of the properties of such representations (some of which were mentioned in the previous chapter and others will be discussed below).<sup>48</sup> In particular it could provide an explanation of many of the apparent metrical properties of mentally-represented space. For example:

- (1) If we represent the fact that A is further from C than from B ( $AC > AB$ ), then there would be a greater quantity of represented space (as distinct from a representation of more space, which makes no commitment about “amount of represented space”) between A and C than between A and B. In other words distance is represented in some form (perhaps in an analogue form) so that each point in empty space is somehow explicitly represented
- (2) If we represent A, B, and C as being ordered and collinear, then there would be an explicit representation of B as being *between* A and C (where by an “explicit representation” I mean that the relation “between” does not need to be inferred, but can be read off in some noninferential means, such as by pattern matching),
- (3) If we represent three objects A, B, and C then it would always be the case that the distance from A to B plus the distance from B to C would never be less than the distance from A to C (i.e., the *triangle inequality* of measure theory would hold so that  $AB + BC \geq AC$ ),
- (4) If we represent three objects A, B, and C so that AB was orthogonal to BC, then for short distances AB, BC and CD it would be the case that  $AC^2 = AB^2 + BC^2$  (i.e., distances would be locally Euclidean so that Pythagoras’ Theorem would hold for short distances).

---

<sup>48</sup> A recent puzzling turn in this discussion is that Kosslyn now claims that the depictive display need not be literally spatial in the way described in the earlier quotation (page 000). He says (Kosslyn, Thompson, & Ganis, 2006, page 131), “The fact that topographically organized areas are physically depictive is irrelevant for present purposes. The neurons in these areas could be interconnected arbitrarily, but as long as *fixed* connections to areas further downstream ‘unscramble’ the activity in earlier areas appropriately, the earlier areas will function to depict.” You can certainly get the same behavior from a display that does not *look* spatial, as long as the relevant spatial properties can be recovered by a mechanism that in effect remaps it back to the literal spatial form. Indeed, you can get it from a representation in English so long as you can use that description to reconstruct a spatial display. There are two problems with this idea. One is that the depictive display is still relevant to the function of the fixed architecture (it’s what the downstream circuits have to “re-map” to), which raises the problems that befell the brain space option (as well as the problems of cognitive penetrability discussed in Section 4.5). The other is that unless the downstream connections actually unscramble the information back into a depictive display, this version of the depiction story no longer sits comfortably with either the phenomenology or the psychophysical properties that made the picture theory attractive, such as the assumption that mental images have metrical properties including distance, visual angle and certain patterns of resolution, and so on. Being able to recover these properties is not the same as having them and it’s the latter that make the pictorialists’ claims principled (it’s the existence of real distance and not some simulacrum that give a principled explanation of such empirical relations as  $time = distance/velocity$ ). I will return to this question in section 5.3.2(2).

(5) If A is represented as being above B and C is represented as being below B, then there would also be an explicit representation of A above C (so 3-term series problems could be solved by “spatial paralogic” , De Soto, London, & Handel, 1965)

(6) If object D is added “far enough away” from the representation of A, B, and C, it would not affect the spatial relations among A, B and C (so “irrelevant properties and relations do not change existing spatial relations”).

The last property (6) is the sort of requirement that in its most general form raises what people in Artificial Intelligence call the *Frame Problem* (Pylyshyn, 1987): In planning a series of actions one must infer all the possible consequences in the entire represented universe for each possible action, including which relations remain invariant, because there is no a priori way to catalog everything that is relevant to a particular action – in principle anything can be relevant to any inference because inferences are holistic (that’s why this is sometimes also called the *relevance problem*). Having a physical spatial layout of the represented situation might solve the frame problem for the properties that were mapped onto the model. Otherwise property (6) requires an inference and may entail “frame axioms” which state what won’t change for various actions.<sup>49</sup>

In addition to these geometrical questions that can be dealt with in a straightforward way if there were a real space in the brain onto which the represented space could be mapped, there are dynamic properties that could be addressed, with the addition of other assumptions. For example, we know that it takes longer to imagine traveling a longer imagined distance than a shorter imagined distance (the so-called image-scanning phenomenon). If a ‘longer distance’ were mapped by a monotonic (or at least locally affine) projection onto longer brain distances, then the equation  $Time = Distance \div Speed$  would apply and would therefore provide an explanation for this regularity while other forms of representation would not (i.e., there is no principle of the form  $Actual-Time = Represented-distance \div Represented-speed$ ). Problems such as these (and others as well) do not arise if space is mapped onto real space.

But is there any reason to think that imagined space is represented by cortical space in this way? Is there any reason to think that when we imagine or think about things laid out in space we create what I called (in Chapter 3) a *neural layout* in the brain? I postpone that question until later since this is one of the constraints addressed by the proposal I will offer.

### 5.3.2 Functional space and principled constraints

#### (1) *Virtual space?*

Suppose, despite its *prima facie* attraction, we find no support for the assumption that perceived space is mapped onto real space in the brain. Can we still reap the explanatory benefits with something less than literal space? Can we, in other words, get some of the benefit of real space in some sort of “virtual” or “functional” space? This is an option that is very frequently raised (see the Kosslyn quotation on page 000). Consider the phenomena you would like the representational system to explain. You would like it to account for the sorts of

---

<sup>49</sup> Notice that even in this simple example I had to add the qualifying phrase “far enough away from” for the example to hold. Otherwise if A had the relation “resting on” to B, and D were added “between” A and B, assumption (6) would fail since A would no longer be “resting on” B. What makes the frame problem so difficult in general is that there is no limit to the relations that objects can enter into and so there is no limit to what might change when a new relation is added to the representation. Consider, for example, the effect on the representation of adding the relations: “Same\_Color(x,y)”, or “Turns\_Green\_If\_within\_2mm (x,y)”, or “Connected\_by\_string(x,y)” If the first and second relations are added to a representation, computing what will change involves checking every pair of objects, of which there might be very many, and will require recomputing colors after any object is moved, including location changes that are inferred from the third relation.

geometrical regularities mentioned byway of examples in the six points on page 000 above (e.g., it should allow noninferential recognition of such things as the relation “between” that holds of the second item when three ordered items are arranged collinearly, that for three imagined items A,B and C, the distance AB plus the distance BC is never less that the distance AC, and so on). Also you would like the sorts of properties that Shepard discussed (under the title “kinematic principles”) to hold. In addition, if it is a Euclidean space (which it should be, at least locally, if it is to be veridical) Pythagoras’s theorem should hold. Simple physical laws such as the relation between time, distance and speed should also hold. This requires a homeomorphic mapping (a continuous neighborhood-preserving mapping) of space onto a set of properties that bear the same relations to one another as do the properties of space itself. Why couldn’t this be done computationally, using a symbolic data structure such as a matrix (as mentioned in the quotation)? This is the most widely-cited option in the literature on spatial mental images (Denis & Kosslyn, 1999).

There is a simple reason why a data structure such as a matrix will not do. The reason is that a matrix is a formalism that *embodies no processing constraints whatever*: It can have whatever properties one stipulates it to have. It can represent empty places or not (sparse matrices in a computer generally conserve space by not keeping empty cells but generating them as needed, much as list processing languages deal with the addition and deletion of cells), it can be traversed by moving through intermediate cells or not (and even the property of being “intermediate” has to be stipulated for each relevant cell), it can be used to exhibit the rotation of an object by retaining its shape rigidly or not, it can take more time to move through more cells, representing greater separations, or not. The mere fact of its being a matrix does not ensure any constraints at all. But then why does a matrix appear to be more natural, compared to representations such as a set of sentences in the language of thought? There is only one reason why operations such as moving a focus of attention or translating or rotating a rigid shape appears to be more natural in a matrix representation. The reason is that since we think of a matrix as being two dimensional, it is natural for us to think of it as a simulation of locations in real space. But if that is true, then it is the real space being simulated that has the properties in a “natural” way – in which case it is the real space being simulated that provides the principled constraints. Constraints, such as the requirement that one move through intermediate cells, is natural only if the matrix is a simulation of space. The matrix itself is only a 2-dimensional object because it is convenient to use pairs of names to characterize individual cells. In fact in the computer the cells are actually accessed by a single atomic name that fuses the two individual names that we consider to be “dimensions”. As for the apparent naturalness of moving through adjacent cells in a computer implementation, this relies on the convention of viewing names as numerals and on the existence of the additional/subtraction operations (and possibly also matrix multiplication) in most computer architectures – none of which is likely to apply to the architecture of the brain.

What is relevant for our purpose in explaining the metrical nature of spatial representation is that a data structure such as a matrix either imposes constraints on possible operations or it doesn’t. If it doesn’t it is of no interest in this discussion because it does not explain anything. If it does, there still remains the additional question of whether it does so because of an implicit assumption that it is simulating real space (which derives its explanatory force from the simulated space, thus making this strategy equivalent to the “brain space” option) or because one merely stipulates that this is how it should behave in order to match the data at hand. None of the relevant space-like properties is an inherent property of the matrix format, they are mere collateral stipulations. But in that case the matrix – or any other “functional space” proposal – is no different from any other form of representation, notwithstanding its intuitive appeal. It is important to see this problem in terms of the question of what constrains the model to have

certain properties. If the constraints must be stated as additional assumptions then the format does no work in explaining how the spatial character comes about since such additional assumptions can be added to any form of representation.

Here is another way to look at this issue. In appealing to functional space to explain many of the spatial properties of thought and imagination we need to ask whether properties we appeal to are fixed properties of the functional space, or are stipulated simply to match the data at hand. This can be viewed as an issue of degrees of freedom: if we postulate ad hoc property P to account for some particular phenomenon, then P serves as a free empirical parameter. We always need *some* such parameters, since we are building a theory in part to account for data at hand. But the more such free empirical parameters there are the weaker the explanatory power of the theory. The goal is to account for the greatest range of phenomena with the fewest such free parameters. How can we reduce the number of free parameters? One way to eliminate free parameters is to determine which parameters are due to the relatively fixed properties of the medium of representation, or the architecture or brain structures, thus decreasing the options available for fitting the given data, thereby increasing the explanatory power of the resulting theory. This is the strategy of attributing as many empirical phenomena as possible to the fixed architecture of the visual (or imaginal) system – a policy that I have strongly advocated (Pylyshyn, 1991, 1996) (see also Chapter 3, section 3.2.2). What, then, if we were to assume that the spatial properties arise from fixed brain properties other than the spatial layout of topographically arranged neurons?

## (2) *Space as a property of the architecture*

Many people feel uneasy with the argument that in order to account for the spatial properties of what is represented in a way that is consistent with the phenomenology, they must postulate a real spatial display as opposed to some functional analogue. To make this uneasiness explicit, consider the following thought experiment in which we start off with a real quantized<sup>50</sup> spatial display in the brain – the sort of “depictive” neural layout postulated in the earlier Kosslyn quotation (page 000), together with associated neural connections appropriate for the next stage in the neural pathway (e.g., some pattern-detection function). Now imagine that the neural fibers are long enough to allow us to move the cells in the early layer (V1) around – to scramble the physical locations of these cells while keeping the connections the same. The result no longer appears to be a real spatial display, yet the function of the network remains the same since the connections have not been altered. This is a case in which the spatial character of the representation appears to derive from the fixed neural wiring and is independent of the geometry of the physical arrangement of cells. Does such a set of cells implement a “functional space” or does it still count as deriving its properties from real space? The answer, I claim, depends on whether the pre-scrambled locations are part of the explanation of the system’s function.

This example is an instance of a class of spatial representations in which the spatial functions arise from the fixed properties of the perceptual or the cognitive system – from its architecture. For purposes of the present discussion I do not make a distinction among various ways that the architecture itself might constrain the behavior. These might include proposals in which spatial properties are imposed by an analogue mechanism of some sort, or any fixed mapping of space onto properties of the cognitive architecture, so long as the right relations among spatial properties is maintained. An example might be one suggested by (Kosslyn, Thompson, & Ganis,

---

<sup>50</sup> The assumption that the display is quantized into a finite number of cells would appear to be unimportant since the cells can be made arbitrarily small. But in fact a quantized display leads to problems if taken literally since distances measured in terms of cells vary with direction, so such a space is not isotropic (and distances do not obey Euclidean axioms). But since the idea of a spatial array has enough problems I will not dwell on the problems raised by quantization of the space into discrete tessellations.

2006) (see also Footnote 49 above) in which the “depictive” array is actually realized by some complex but fixed network of neurons which treats certain pairs of neurons as representing adjacent places even though the neurons themselves may in fact be located far from one another in the brain. This option uses what might be called an encrypted version of literal space. The question is: Does this mechanism use a literal space or a functional space, according to the present taxonomy?

To count as literal space, as I understand the term, the essential requirement is that the system must get its explanatory force from real spatial properties, so that the sample of properties listed at the beginning of section 5.3.1 must hold. The system of representations counts as using literal space if it meets two requirements: (1) The properties derive from a fixed architecture and therefore are cognitively impenetrable, and (2) One must appeal to real spatial properties in providing explanations of why certain behaviors hold. For this to be true of the encrypted space option, the real spatial layout must not only be recoverable (decryptable) but it must be cited in providing explanations because it is the source of the principles that determine the behavior of the neural circuits. In other words the circuits do not have the properties they do by mere ad hoc stipulation (they are not free parameters) but because they really do derive from spatial locations and it's these locations that explains their geometrical properties. For example, some pairs of cells are treated as being adjacent to each other while other pairs of cells are treated as being far from one another. Which pairs count as close together and which count as far apart cannot be determined without the decryption function. Without the recoverability of a spatial display there would be no independent motivation for the spatial properties that the representation exhibits. Thus the answer to a question such as; Why does it take longer to switch attention a greater represented distance? refers to the literal meaning of “distance” in the decrypted display, not to properties of architectural network onto which distance is mapped. That's because the equation  $t=d/v$  has the explanatory force of a nomological principle whereas any other ad hoc arrangement of neurons that happens to yield this same pattern begs the question; Why that pattern rather than some other? It is no accident that talk about the nature of spatial representations makes essential use of spatial terms (e.g., patterns in representations are said to be bigger, above, inside, etc., in relation to other patterns). The frequently-cited notion of an analogue architecture is a special case of a fixed architecture, though perhaps with some added requirements, depending on your understanding of what constitutes an analogue representation.<sup>51</sup>

What, then, do we say about the encrypted-space option, or any option that attributes spatial properties to the architecture of the visual (or imaginal) system? Such architecture-based explanations do not circumvent the most serious criticisms raised in Chapter 4 in connection with the literal brain-space proposal.

(a) Architecture-based explanations for how spatial properties arise are not consistent with relevant empirical data. By definition, properties of the architecture do not change in ways attributable to changes in beliefs – they are cognitively impenetrable. Thus empirical phenomena ought to be insensitive to beliefs (e.g., about the task or about how events would

---

<sup>51</sup> Despite the existence of clear and easily understood cases of analogue processes, and despite the frequent references made to this notion, it remains poorly understood. In particular it has turned out to be extremely difficult to give an acceptable set of conditions for something falling under that category. I have used the term to refer to any case where behavior is attributable to a fixed architecture (e.g., Pylyshyn, 1984, chapter 7). Others have reserved the term for processes or representations involving continuous properties (see, for example, Goodman, 1968; Lewis, 1971). Jerry Fodor (Fodor, 2004) has recently argued that what makes something an analogue representation is that it doesn't have a canonical decomposition into semantically-interpreted constituents – in fact it doesn't have constituents at all. Whether one calls a neural layout an analogue representation or not, states of this sort, which may be used for clustering and establishing correspondence for apparent motion, are a type of nonconceptual quasi-representational information-carrying state.

unfold in the world). But the data do not support this assumption. Consider any empirical phenomena described in terms of size or distance, such as observed with different mental size or different represented distances, such as mental scanning experiments. These phenomena are cognitively penetrable (see Section 4.5.1). So far as I am aware all examples that are described in terms of spatial features of mental representations are either penetrable or their robustness is due to their task demands – i.e., subjects take them to require that they simulate what they would see if they were to witness the relevant event. As I pointed out in note 48, if you are require that the task be accompanied by the specified conscious content (e.g., “imagine that the mouse occupies a small part of your field of view”, or “find the named place on your image of the map”) then this requires that you recreate the experience of seeing the event unfold and that, in turn, depends on your knowing what things would look like, how long they would take, and so on, which adds another constraint to the resulting observed behavior. The way things *would look* is the way you believe they *would be*, hence the conformity of the data to the theory that postulates a spatial display.

(b) Architecture-based explanations are not compatible with the natural and intuitively satisfying story that those who appeal to a literal-spatial display tell about why, for instance, it takes more time to scan greater imagined distances or why it takes longer to detect small features in a small image. Recall that the story for the former, which agrees with the phenomenology that motivates the literal display account, is that it takes longer to scan attention between imagined places that are further apart on the original scene because the distance on the representation is greater (and you have to scan over empty spaces in the representation). Without that story, which refers explicitly to image distance, the attraction and motivation behind the display view is lost. Similarly in the image-size phenomenon, it takes longer to report details from a smaller image because fine details are harder to see and so one may have to “zoom in” on the image first. In each case a literal spatial property (in these examples, distance or size) has to be cited. If we attribute the time function to the architecture without mentioning any spatial properties, the claim becomes that, in effect, it takes longer because that's just the way it is given the kind of brain we have; and that claim has to be stated without adverting to distance or size on the representation. There is nothing left of "distance on the image" which is the reason for the interest in the mental scanning experiments (and is why they have served as a “window on the mind”, Denis & Kosslyn, 1999). Consequently we are left with no principled *explanation* of the scanning results or the image-size results, only a restating of the finding.

To summarize, the issue here is simply that an explanation of spatial properties needs to appeal to principled, as opposed to ad hoc properties (or post hoc stipulations). A “functional space” may or may not be principled. It would be principled if it was viewed as a *simulation* of real brain space since in that case it's the real space (in the brain) that provides the explanatory principle and the matrix is simply a convenient way to implement it in a conventional computer (otherwise you would need a computer with a display and a visual system to examine it, etc.). Of course this version is simply the picture-in-the-head alternative which has many other problems (see section 4.5.3). The architecture-based account could also be principled if it was the result of independently motivated properties of the architecture, but in the case of the spatial examples, such as the distance and size phenomena, the independent motivation must be based on properties of space so literal spatial properties are not avoided in this way.

### **5.3.3 Internalizing by incorporating visuomotor experience: Poincaré's insights**

There is another way to understand our sense of space which approaches being an internalizing view, but only insofar as it emphasizes an internal link with the motor system. In

what follows I will make much of this general idea and it will lead to the notion that rather than internalizing space, the converse actually holds: The mind actually externalizes space by projecting spatial relations onto visual-motor and proprioception-based sensory information. The basic idea for this direction can be traced to the work of Henri Poincaré early in the 20<sup>th</sup> century.

In a series of essays written almost a century ago, Poincaré analyzed the nature of space. In one of these essays he describes how a three-dimensional sense of space might arise in a sentient organism confronted with information in various forms and modalities and in many (not just 3) sensory dimensions (Poincaré, 1963/1913). A central idea in Poincaré's account is that the organism has to be able to distinguish between experiences that correspond to changes in position and those that do not.<sup>52</sup> According to Poincaré, the key to being able to recognize the difference depends on being able to distinguish between changes brought about by our own actions and changes that are externally caused. Here Poincaré makes use of the notion of the reversibility of certain sensory states – what he calls a “correction” – whereby we can undo an externally caused change by a complementary voluntary change that brings back the original sensory state. Suppose, for example, that you are touching some object with the index finger of your right hand and receiving the tactile signal *T* while at the same time you sense the object visually, yielding the visual signal *V*. According to Poincaré, if the object moves from external causes, you will perceive the visual signal change from *V* to *V'* and the tactile signal will disappear. But you may be able to bring back the original tactile signal *T* and visual signal *V* by an action, represented by a series of muscular movements accompanied by a sequence of proprioceptive sensory signals  $S_i = (S', S'' \dots)$ . This action in effect undoes the movement of the object in space by a compensatory voluntary movement. Moreover, this same “renewal” of the tactile signal can be accomplished equally by any member of an equivalence class of sequences  $\{S_1, S_2, S_3, \dots\}$ . What the members of this equivalence class have in common is that they can be described as “moving your finger from a common starting position to a common final position”. According to Poincaré what you, or your evolutionary ancestors, have internalized is the principle that if you are touching an object and your visual signal changes from *V* to *V'*, you can once again touch the object by carrying out a motor sequence in which the proprioceptive signal follow the sequence corresponding one of the members of the equivalence class. Thus the basis for your knowledge of spatial locations is this skill of moving in such a way as to bring back a tactile or visual signals (as noted in footnote 3 my use of the terms “sensation” or “sensory signal” refers to sensory information and does not assume that they are conscious).

Poincaré used the notion of an equivalence class of sequences of sensory signals that correspond to moving a finger from a particular initial position to a particular final position as a way of defining a common location across the several fingers and hands. The classes of movements define “spaces” and the spaces marked out by each finger are then merged by the recognition that when two fingers touch one another they define the notion of “same place” and so lead to the convergence of the initially distinct spaces. Poincaré then goes on to argue that the reason that our representation of space has 3 dimensions, rather than 2 or 4 dimensions, the way that the equivalence classes are established, together with the fundamental utilitarian boundary condition that we should not count as equivalent two sequences of sensations that fail to take us to the same final position (where the tactile sensation is renewed) nor should we count as equivalent two sequences of sensations that take us to different final positions (where the tactile sensation is not renewed). It is these boundary conditions that force the tri-dimensionality of

---

<sup>52</sup> Interestingly, Austen Clark (Clark, 2000) makes a similar point when he discusses reasons why sensations that correspond to different locations are fundamentally of a different kind from sensations that corresponding to qualitative properties such as color, size, shape and so on.

space. The reason I have belabored this point is that apart from providing an elegant account of the basis for the dimensionality of space, Poincaré's analysis touches on several issues that will be relevant to our present discussion.<sup>53</sup>

The details of this analysis don't carry much conviction these days, and indeed the reversibility of sensation condition was criticized by Jean Nicod, but many of the ideas remain sound. For example, the first point of contact between Poincaré's analysis and the ones I will propose concerns the recognition that there are two distinct kinds of changes in sensory states; those that signal a difference in location and those that signal a difference in some sensory quality, say color or texture. Whether or not you like Poincaré's way of making the distinction, in terms of the capacity to "correct" or revert to an earlier location-marking sensory state<sup>54</sup>, the distinction does play an important role in recent discussions of sentience, and is especially central in the work of Austen Clark (Clark, 2000), though for different reasons. The second point of contact concerns the emphasis placed on sequences of muscular actions and sensory inputs and to equivalence classes of such sequences. This a remarkably modern idea, although it is not expressed in this way in current writings. What Poincaré's analysis shares with contemporary analyses of what I called the "sense of space" is the idea that the nonvisual apprehension of space may be a construct based on mechanisms that compute the equivalences among otherwise very different sequences of muscular gestures. Computing the relations among representations of positions of limbs, sensors, and other movable parts of the body is arguably one of most ubiquitous and best understood functions of the brain – functions carried out primarily in the posterior parietal cortex, but also in the superior colliculus, in the motor and premotor cortical areas (for a review of the biological basis of this function, called a "coordinate transformation", see Gallistel, 1999).

Computing one position-representation given a different position-representation is commonly referred to as *coordinate transformation* (CT). One way to view CTs is as a function from the representation of an articulated part of the body (e.g., the eye in its orbit) in one orientation to the representation of that part (or a different part) in a different orientation or relative to a different frame of reference. It also applies to computing a representation of a location within the reference frame of one modality to a corresponding representation in the reference frame of another modality. The relevant representations of limbs in these cases are typically expressed within a framework that is local to the parts in question – such as the states of the muscles that control the movements, or the joint angles that characterize their relative positions, or to endpoint locations relative to the body. The relevant representations of sensory inputs may similarly be in proximal coordinates (e.g., locations on the retina or on the basilar membrane) or other local coordinates.

---

<sup>53</sup> Poincaré's examples use fingers and the capacity to sense the locations of fingers. His essay was on my mind at the time I was formulating the FINST Index theory and may be the reason for the appearance of "finger" in FINST.

<sup>54</sup> This use of reversibility of sensory states as the signature of voluntary movement plays a central role in a recent model of how properties of space can be inferred from minimal information about the dependencies between actions and sensory inputs (Philipona, O'Regan, & Nadal, 2003). There is in fact a great deal in common between the present analysis and the ideas on the importance of sensory-motor factors in visual perception proposed by the Kevin O'Regan and his colleagues (and spelled out in O'Regan & Noë, 2002), although some of the claims made in the latter publication on the nature of conscious perception seem to me unconvincing (for reasons I allude to in my commentary on the O'Regan & Noë paper, in Pylyshyn, 2002b).



The importance of these ideas in the present context relates directly to the theme of nonconceptual contact between mind and the world. In particular, since I have argued (in Chapter 3) that this contact does not begin with the selection of spatiotemporal regions I need to say how places in space are represented – and indeed whether they are represented *as such*. What I will do in the last part of this chapter is to consider another approach to the question of what it means for the nonconceptual visual system to index or pick out a place or region in space. We have already discussed the problems with the traditional view that the first, nonconceptual (or sentient) contact with the world occurs through the detection of features-at-locations (the idea developed by Peter Strawson as part of his analysis of the representation of particulars, and adopted recently by Austen Clark in his theory of sentience). What I want to do now is suggest another way in which what appears at first glance to be spatial selection might be achieved without the selection of places, specified in terms of a unitary frame of reference. But first I will step back to see if some general conditions on this problem can be set out.

## 5.4 What is special about representing space?

### 5.4.1 Conditions on a theory of Active Spatial Representation (ASPAR)

The impression that spatial representation is different from other kinds of representation is usually associated with examples from perception and spatial reasoning (or mental imagery). In other words the evidence comes from phenomena that involve active working memory, as opposed to more permanent long-term-memory. There is reason to think that the representation of space under these conditions, which I will refer to as *Active Spatial Representation* or ASPAR is special in a number of ways.<sup>55</sup> Although many of the characteristics of ASPARs are also shared by other nonspatial representations, representing space appears to involve all the following characteristics in essential ways.

#### (1) ASPARs must be able to represent magnitudes in a continuous manner

Psychophysical evidence shows that we have encodings of relative magnitudes and that the magnitudes that are encoded have a particular systematic effect in reasoning. Examples of such systematic effects include *scalar variability*, in which error variance scales with the mean (so that the coefficient of variation,  $Variance(T)/T$ , remains constant), *Fechner's law* (or the Weber-Fechner Law), in which the discriminability of the magnitude of two properties varies in proportion to the mean magnitude of the properties (i.e.,  $\Delta T/T$  is constant), and the *symbolic distance effect*, in which the time to compare two magnitudes, including non-sensory ones like cost, increases in proportion to the ratio of the magnitude of the difference to the mean magnitude, as in Fechner's Law. Although the inference from psychophysical regularities to conclusions about the format of representations is by no means free of problems, since it requires some (unspecified) assumptions about the nature of the mechanisms that interpret the representations, it has led many people to conclude that at least some magnitudes may be represented in analogue form, where both the content and the vehicle of the representation are continuous (see note 52).

---

<sup>55</sup> While I am loath to propose yet another acronym, the closest constructs in general use, “visual short-term memory” or “visual workspace”, both have the wrong connotations since what I am trying to capture is not really a type of memory, visual or otherwise, or a storage place, but rather the representation one constructs when reasoning about spatial properties and relations. I make no claims about whether this is located in a distinct place in the brain, whether or not it is part of a more general amodal memory, and whether or not it is conscious.

*(2) ASPARs must represent stable spatial configurations*

Spatial configurations involve relations over multiple objects – in that sense they are holistic and require simultaneous access to multiple represented objects. Access to such configurational information may allow some spatial consequences to be arrived at by pattern match or by table lookup, without inferences based on independent geometrical axioms. Take for example, the well-known three-term series problems which are assumed to involve spatial mental images (Huttenlocher, 1968). These are problems involving ordering syllogisms such as: “John is taller than Mary; John is shorter than Fred; who is tallest (or shortest)?”. The explanation of many of the observed effects is that the subject forms a spatial image, using an imagined spatial relation such as “above” to map “taller”, and then is able to “read off” the correct answer from the resulting spatial array without apparently using the axiom of transitivity or rule of syllogistic logic (of course such a rule is implicit from the very start since it determines which spatial property in the representing medium is of the same logical type as the relation in the problem being represented; the relation “above” could not be used to map an intransitive symmetrical relation such as “is married to”). Such “reading off” the properties of a spatial display requires that configurational or pattern information among all relevant objects be available at once, or at least that they remain in place as the objects are examined.

*(3) ASPARs must represent spatial relations in 3D*

Relations in depth must be in the same format as the encoding of relations in the plane since the two must work together seamlessly. Experimental evidence from such phenomena as ‘mental rotation’ or ‘mental scanning’ show identical functions in depth as in the plane.<sup>56</sup> As we saw in the last chapter, this 3D requirement is one of the decisive reasons why a “mental image” cannot be identified with a topographical display in visual cortex.

*(4) ASPARs represent spatial properties acquired through different modalities*

It seems that at least some spatial representations may be amodal, or multimodal across the auditory, tactile and proprioceptive modalities (Spence & Driver, 2004). Patterns and locations perceived initially in one modality are often remapped onto other modalities. For example, Jon Driver and Charles Spence carried out a number of ingenious experiments that involved presenting cues in one modality and observing the effect on discrimination tasks at the same general location in another modality (described in Driver & Spence, 1998, 2004). They found a high degree of cross-over, so that a cue on the left side of a visual display improved elevation discrimination (which of two signals is higher?) of both auditory and tactile stimuli on the cued side. Interestingly, with tactile stimuli the same result is observed even if the hands are crossed, showing that it is the location of the tactile stimuli in extrapersonal space, rather than their location relative to the body that matters. Similarly, when subjects expected a sound on one side, both auditory discrimination and visual judgments on that side are improved – even if visual events were twice as likely to appear on the other side. Perhaps even more surprising, in an experiment in which subjects had to judge relevant tactile stimuli on one hand, while ignoring concurrent tactile distractors on the other, the distractors interfered less when the two hands were held further apart – so again it seems that it is the location of tactile stimuli in external space that matters even when the task is to discriminate relative location on the skin. But spatial location information is not entirely shared across modalities. There have been some reports that we

---

<sup>56</sup> An example of the “mental scanning” experiment is briefly sketched below (Section 5.5.2). The “mental rotation” experiments are not described (but see, Pylyshyn, 1979, 2003). This refers to the finding that when asked whether two oriented shapes are congruent or mirror images (enantiomorphs) of one another, the time it takes is proportional to the angle between them – in either 2-dimensions or in 3-dimensions (the original 3D finding is reported in Shepard & Metzler, 1971).

cannot easily recognize shapes if some of the location information is presented visual and other information is presented in the tactile or auditory modality. The question whether one can transfer shape information across modalities without learning is called “Molyneux’s problem” and it appears that very young infants are able to transfer at least simple shapes (Meltzoff, 1993), so cross-modal shape representation may be at least partly an innate capacity.

The unitary first-person experience of space, together with demonstrations of cross-modal spatial attention has promoted the assumption that we have a *single global representation* of space in an allocentric frame of reference. I will suggest shortly that this assumption is not warranted if it is interpreted to mean there is a single representation of space *in a single frame of reference* that cuts across perceptual modalities as well as thought and imagination.

(5) *ASPARs must be able to engage the motor system*

One of the characteristics of “spatial representations” is that we can “point to” represented things (e.g., in our mental image). When you imagine your office you can think (and even make corresponding pointing gestures while having the thoughts) “my computer screen is *here* and my keyboard is *there*”. Eye movements when examining a mental image have a similar pattern to those observed in vision (Brandt & Stark, 1997). One of the things that makes mental images seem spatial is the ability to orient to and refer to parts of the image in a demonstrative manner (using the terms “here” and “there”) in referring to objects in one’s mental image. But there are also significant differences in detail between motor reactions to mental objects and to actual perceptual ones. While we may gesture towards imagined objects, they do not engage the visual-motor system the way perceived objects do. Detailed analyses of gestures such as reaching and grasping show that these actions differ in detail when the object of the reaching is remembered or imagined than when it is seen. Actions toward remembered objects have the characteristics of pantomimed movements rather than movements that are under the control of vision (Goodale, Jacobson, & Keillor, 1994). This suggests that visual and imagined or recalled objects may engage different perceptual-motor systems (Goodale & Milner, 2004).

(6) *ASPARs must capture the continuity and connectedness of space*

This is an important ingredient of spatial representation even though it is not clear exactly what it entails. In particular it leaves open questions such as the following (the proposal I will be making later does provide tentative answers to these questions):

- Are empty places explicitly represented *as such* (as opposed to there being some indirect way of answering the question, “is there something at location x”)?
- Does the system of representation (or the architecture) itself ensure that distances meet **metrical axioms** (e.g., the triangle inequality  $AB + BC \geq AC$ ) or that they are **Euclidean** (e.g., follow Pythagoras’ theorem)? Or is that the result of subjects’ tacit knowledge of geometry?
- Are representations of the movements of objects (or the movement of representations of objects) constrained so that in getting from A to B they must pass through ‘*intermediate*’ locations? And must they take longer if the distance from A to B is represented as being greater?
- Are primitive non-conceptually-mediated perceptions of such properties as causality a result of properties of the ASPAR mechanisms, or do they have their origins elsewhere? For example is the perception of causality in the Michotte tunnel effect or in the launch effect (where a moving object collides with a stationary one which then begins to move, and is perceived as having been launched) a result of the same mechanism in the ASPAR that enforces the metrical properties of space? Insofar as such effects may be

nonconceptual and originate within the architecture of the visual system, an argument could be made that they are part of what an ASPAR provides non-inferentially. A similar argument has in fact been made by John Campbell who takes the position that the connectedness of space is a matter of the causal structures of space.<sup>57</sup>

These 6 properties constitute strong constraints on any theory of the representation of space in active memory. They clearly are not met – at least not without significant additional assumptions – by a system that represents spatial properties in terms of Cartesian coordinates. They are also not met by the proposal that spatial properties are represented by locations in a “functional space” since that begs the question of why the functional space has the assumed properties, and they are not met by the proposal that spatial locations are mapped onto locations on the surface of the visual cortex – though in that case it fails for a number of purely empirical reasons. Before I present a proposal that offers a provisional answer to these questions I will briefly touch on one special type of spatial representation and ask whether it might use the same form of representation as used in ASPARs – this is the spatial representation involved in navigation (i.e., “an internal map”).

#### 5.4.2 Maps and map-like navigation behavior

I have been focusing exclusively on spatial representations constructed in perception (primarily vision) and thought, which I refer to as Active Spatial Representations (ASPARs). Although this remains my focus in this chapter, largely because it provides an interesting application of the ideas introduced in earlier chapters, I should at least mention one other context that involves rich spatial information, that of navigation. Because navigational planning is itself a task that involves reasoning about spatial layouts, it is plausible that navigation may involve a form of representation similar to ASPARs. In the context of navigation the notion of a map-like representation or of “map-like behavior” (e.g., Menzel, Greggers, Smith, Berger, Brandt et al., 2005) has been widely invoked since Tolman’s introduction of the idea of a “Cognitive Map” beginning in the 1920s (see, e.g., Tolman, 1948). When engaged in finding their way from a starting location to a goal, or between geographical landmarks, people often navigate by using a 2-dimensional display (the usual sort of printed terrestrial, marine or road map) which represents the relative locations and the types of various visible landmarks, and the nature of the terrain through which the person will travel (together with a scale so distances actually can be measured on the map). Similarly when navigating from memory, one’s experience is often that of examining a map with one’s “mind’s eye”, which suggests the picture-in-the-head view that I have been discussing here and in the previous chapter (as illustrated in Figure 5-2). Partly for this reason the notion of cognitive map has frequently been associated with a 2-dimensional display such as what many people have assumed constitutes an ASPAR.

The salient empirical fact is that for many animals, as well as for birds and insects, the spatial information available for navigation appears to be not only quantitative and remarkably precise but also allows the animals to engage in behavior that looks very much like the product of goal-directed planning. The complexity of this behavior leads to the assumption that these animals (including ants and bees) have an internal “map” of some kind. The problem is, as it was in the case of reasoning with the aid of a mental image, to specify not only the content of map-like representation, but also the constraints imposed by the map format. Here the research with rats,

---

<sup>57</sup> In (Campbell, 1993, p69) he writes, “It is only its figuring in an ‘intuitive physics’ of one’s environment, through regularities connecting spatial properties with other physical properties, that makes it reasoning that is not purely mathematical, but rather, reasoning about the space in which one lives.” Christopher Peacock (Peacocke, 1993) also takes the position that there is an intimate relation between spatial reasoning and what he calls “intuitive mechanics” but his primary concern is in the question of what is needed in order to have a conception of the material world, not what is entailed by one’s sense of space..

birds and insects may be extremely useful because these organisms have very well developed navigational skills, yet presumably have limited resources and limited capacity to reason from general knowledge, so their behavior may more directly exhibit the content and the constraints of such map-like representations.

Whether ASPARs and navigational maps use the same form of representation is an open question. On the face of it there are some clear differences. Navigation is an interactive process in which perception of the environment plays an essential role as the organism moves through it. Moreover the way in which spatial information comes to be in the “map” is different from the way it comes to be in the ASPAR. In the examples discussed earlier involving ASPARs, the information comes either from perception or it is constructed from cognitive descriptions. By contrast, spatial information in a navigational map is the result of precise measurements of time, distance and direction that are the result of having traveled through the terrain. Insects use a sophisticated method for establishing how far and in what direction they have traveled, The method, called “path integration” requires integrating the vector speed with respect to time, a process that yields a continual record of distance and direction from starting point (Collett & Collett, 2000). To measure speed, bees use motion over the retinal image (visual flow fields) while ants use a measure based on motor output, and other animals may use some other measure or combination of measures (Collett & Collett, 2002). Birds and bees also spend some time doing what might be called reconnoitering the territory through which future travel will take place and in some cases (e.g., in the case of bees) it involves communicating the relevant information (e.g., to other members of the hive). How this information is assimilated from the exploration is largely unknown, but the ability to integrate new measurements during trips with related information stored in memory looks very much like an application of vector algebra since computing the distance and direction between two places that had been visited but where the path between them had never been traveled requires taking vector differences. Similarly, finding a short-cut through unfamiliar territory, or finding a way around a barrier or a dead-end path suggests that the animals can coordinate representations at several scales. It is thus plausible that the map-like information which is stored in long-term memory, and which controls goal-directed movement, uses a different mechanism and form of representation, at least some of the time, than used in reasoning about spatial layouts.

The representations used in navigation also differs in many ways from how spatial information is represented in the usual sort of canonical maps. Local features of landmarks may or may not be represented and may be encoded in a context-dependent way (e.g., ants appear to recognize landmarks only when they are traveling towards the same goal as they were when the landmark was first encoded Collett & Collett, 2002). Also some landmarks may be selected temporarily during travel without encoding their properties. Given the story I have been unfolding in this book, it would come as no surprise to find that when an animal sees a layout of objects it is moving through, it might keep FINST-like pointers on the salient landmarks without encoding their properties or locations, and it might even use its perceptual-motor skill to set a course in relation to these FINSTed target landmarks.<sup>58</sup>

---

<sup>58</sup> I don't want to overstate how much of the planning and course-setting can be done “on line” while traveling a route and tracking landmarks using indexes since (Gallistel & Cramer, 1996) showed that a vervet monkey can plan a minimum distance route that involves looking ahead at least 3 landmarks. While it is not known exactly what information the monkey has stored and what its plan takes into account in advance of initiating its travels, their findings do show that the monkey does better than a nearest-neighbor search. But it is also clear that in executing a plan the monkey must select and keep track of some of the individual target landmarks and final goal at some point in the course of its travel (as in other examples of “visual routine” problems discussed in Section 1.4.2). Since there are reports that monkeys can track at least 2 moving targets among 4 objects in an MOT paradigm (tracking in the macaque has been demonstrated by Mitchell, Sundberg, & Reynolds, 2006, but these investigators did not explore the limits of how many could be tracked), it is at least possible that the plan may refer to individual visible landmarks indexed at the time of initial observation of the landmarks being baited and tracked while executing the initial segment of the plan.

The assumption that properties of selected landmarks may not always be encoded is in fact supported by data from rats and other animals (including human infants). For example, (Cheng, 1986; Gallistel, 1990) showed that a “map” representation may not contain any identifying information for the individual landmarks – only their geometrical relation to other landmarks (and to the larger room space). Consequently, if the information takes the form of a map, it would in this case be a map with no labels or legend. This entails that certain systematic ambiguities in locating objects necessarily results from the representation. For example if a particular goal landmark is located as in Figure 5-1 below, would show no preference between the actual location of the goal G and the mirror-image location G' which bears the same local-geometry to landmarks A, B, C, and D. This is exactly what Ken Cheng found. In fact he found that an animal not only does not represent features of the landmarks, it is unable to take account of the relevant (disambiguating) intrinsic properties (including color, texture, size, shape, odor) of the landmarks (or of the walls AB, BD, DC, CA when these were distinctively marked). In an animal that has been disoriented by being rotated a few times, local distinctive features don't seem to be used in navigating to the food – perhaps local cues have not even been entered into memory and associated with corresponding landmarks (just as we found that properties of objects in MOT were not encoded and used in tracking). These properties can, however, be discriminated and used for other purposes, such as choosing between two targets located side-by-side. This appears to be a quite general phenomenon and is found in human infants as well (see, Hermer & Spelke, 1996). Infants, it seems, also do not use local cues in navigating, just as in the experiment by (Tremoulet, Leslie, & Hall, 2000) described in Section 2.5, were able to use color cues to determine that there were several distinct objects but were not able to use color to recognize whether particular colored items placed behind a screen had changed.



Figure 0-1 Mirror image indeterminacy in the navigation module (Cheng, 1986; Gallistel, 1990)

There are other ways in which animal navigation departs from the ideal that would be expected if the animal were using a canonical map. For example, (Wehner, Boyer, Loertscher, Sommer, & Menzi, 2006) showed that while the desert ant (which has exceptional navigation skills) uses information about some familiar landmarks in navigating to and from its nest (in addition to its principle method of relying on dead-reckoning based on path integration), the recognition of these landmarks depends on whether the ant is on its way away from or back to the nest. On the way out it freely uses landmarks acquired on its previous outward journeys but ignores familiar landmarks acquired during previous inbound journeys and vice versa. This was shown by observing the ants' behavior when an inbound ant is captured and displaced to a point on its habitual outbound path. Under that condition the ant ignores the landmarks and engages in systematic search, repeatedly passing outbound landmarks without recognizing them until it finds an inbound landmark which then allows it to get back to its nest. So the ants' “map” appears to have some feature-based labels but in a way that only applies to one direction of travel.

It would be useful to have an analysis of the behavioral capacities underlying the use of what is generally referred to as ideal “map-like behavior.” For Tolman, being able to take novel short-cuts to the goal was the most important criterion of map use (this includes detouring around new obstacles and ignoring local cues in favor of more global direction in making route choices). This is probably still the most widely accepted criterion. Some investigators have listed additional criteria. For example, (Menzel et al., 2005) cite as symptomatic of map-like behavior the ability to make different choices about where to go next in the course of the travel, even when the animal has been transported by the experimenter to an unfamiliar location. But however impressive these behaviors may be, and even when they tell you something about the content of the representation (what information must have been encoded) they tell you little about the form of the representation involved that makes it map-like.

One attempt to characterize the content of the map-like navigation skill is provided by (Gallistel, 1990), who argues that to exhibit map-like skills is to possess and be able to operate on two location vectors for each represented object in the environment: one vector that locates the object relative to an allocentric frame (e.g., relative to a fixed landmark) and another vector locating it relative to the organism (the egocentric vector). In addition, the organism needs to be able to compute various functions over vectors, such as vector addition and subtraction. This serves as a theory of the information the organism must have and represents an important advance in understanding the process of navigation in animals. It is a competence theory, or perhaps closer to what Marr (Marr, 1977) called a “computational theory” by which he meant a theory of the function that is computed. For example while vectors from an origin to individual landmarks may be sufficient to allow locating each landmark, it may also be that the distance and bearing of some landmarks in relation to other reference landmarks might be stored as well, along with features of the landmarks themselves. Such tradeoffs between processing and memory are well-known in computer science. Many such questions cannot be addressed without a proposal for the architecture of the computational system, for example questions such as the tradeoffs between storing feature information as opposed to location information, the frame of reference in which the location information is stored, and how the information is accessed and turned into route choices. Even a question such as which algorithm is used cannot be answered without a theory of the architecture (algorithms can only be specified relative to architectures, see Pylyshyn, 1984). A theory of the competence entailed by map-like behavior also does not address the question of how ongoing perception interacts with stored map information (for example, how indexes might be used in navigation), and questions about how much of a trip is planned in advance and how reactive the plans are.<sup>59</sup> Also even if the postulated vector capacity describes sufficient conditions for many map-like navigation behaviors, they may not be necessary conditions. Many organisms (probably including humans) achieve what would be called map-like navigation without such a precise computational system with a dual-vector representation. For example, there is some reason to think that we have information at various scales and that information in a local frame of reference has to be integrated with more sketchy information in a more global frame of reference.

This and other issues concerning the representation of map-like information for navigation leave open many questions about the form and content of the spatial representation. For example, they leave open such questions as whether and how the format of the representation yields such properties as those listed in section 5.4.1. This leads to such further questions as

---

<sup>59</sup>The study of automatic plan generation in artificial intelligence has come a long way from the early systems that generated complete plans in advance and has moved to more “reactive” plans that make allowances for what will be discovered in the initial steps of executing a plan (see the discussion of reactive planning in Kabanza, Barbeau, & St-Denis, 1997). The question of just how reactive animals’ navigational plans are is an interesting open problem.

whether empty places are explicitly represented, how magnitudes are represented, whether the information stored in the representation includes features of landmarks, and if so under what conditions are they accessible (e.g., whether these features be accessed freely or only when the actual landmark is attended), whether the medium or format of the map representation constrains distances so that the axioms of measure theory or Euclidian properties hold at least approximately, and whether there is a single global map or a network of local maps stitched together by special computational mechanisms. Notice that these are the same questions raised about ASPARs, which suggests that at least some properties of maps may involve the same representations as used in ASPARs. In this book I do not address most of these questions because I confine myself to the problem of how (and whether) we construct an internal representation of space in the course of thinking about spatial problems (i.e., spatial imagery, whether or not it is conscious) and also to some extent during perception (insofar as I address the picture-in-the-head assumption illustrated in Figure 4-3, that sometimes makes its way into perception as well as mental imagery).

In what follows, then, I will sketch an alternative way to look at the problem of how space is represented in an ASPAR (if it is represented at all) so as to satisfy the desiderata sketched so far. It is a strong form of what philosophers will recognize as an externalist position – though it is not concerned with justification of beliefs, which is where externalism most often appears in epistemology. The view is just this: In imagining a spatial layout visual indexes (FINSTs) are used to pick out concurrently perceived objects that are roughly in the correct relative locations in the perceived world. Each indexed object is associated with a unique label of a recalled or imagined object. These labels allow the system to keep the individual indexed objects distinctive and also allow the visual system to treat the indexed objects as though they were marked – the visual system can thus detect patterns among indexed objects. The spatial properties that concern the mental objects (i.e. the conditions listed earlier) are the consequence of actual perception of the spatial relations among these indexed objects. This simple idea, called the Projection Hypothesis, is developed in the next section.

## 5.5 Externalizing spatial properties: The projection hypothesis

### 5.5.1 Where is our spatial representation?

The standard view about spatial representations is that they consist in some sort of picture or map laid out in the head – most often they are thought to be in the topographic (even retinotopic) areas of the visual system, notably the earliest parts of the visual cortex (area V1). Clearly *something* is in the head when you imagine a spatial layout. But it does not follow that the spatial properties (especially the metrical and geometrical properties) that are available to you derive from whatever it is that is in your head. Even less does it follow that your representation is in a retinocentric or brain-centric frame of reference. The notion that spatial properties are in your head is not consistent with either the phenomenology or the psychophysical data. When we imagine, say, a simple geometrical pattern, the figure may feel like a picture, but it is also clear that it is imagined as being *out there*, and not in the head. When you move your eyes, turn your head or turn your body, the image does not naturally move but remains fixed in extra-personal space, in what is called the allocentric frame of reference. As a result it must move in the retinotopic frame of reference and it takes extra effort to imagine it as fixed on your retina as you turn or move your eyes. But you are typically unaware of making eye movements while you imagine, and so it would be odd if you experienced the image as moving with your saccades (just as it feels odd to experience an afterimage because that *does* move with eye movements). In fact (Brandt & Stark, 1997) have shown that you explore your mental image using eye movements that are very like the ones you use when you examine an actual scene, which provides even



more reason to think that you locate your image in extra-personal space. Even if you walk towards or around the object you are imagining, it remains fixed in space. Do the following experiment. Close your eyes and point to some places in the room you are in. You will find that you can do so with a fair degree of accuracy. Now imagine yourself standing in a different part of the room – say on the opposite side of the room and repeat the exercise by pointing as though you were at this imagined location. You will find that your accuracy in pointing to the same places drops considerably. Now if instead of imagining yourself on the other side of the room, you close your eyes and *walk* (or are led, blindfolded) to the spot on the opposite side of the room, you will once again find that you can point to places with nearly the same accuracy as you did when you were in the initial position. And, as you might expect, you are now very bad at pointing to where things would be if you were still at your starting position, or if you imagine that the things in the room moved with you as you walked (these findings have been demonstrated by a number of investigators, see e.g., Farrell & Thomson, 1998; Farrell & Thomson, 1999). Real movement interacts with imagined objects in a very natural way and such movements treat the imagined objects as fixed in space (so that, for example, your relation to your mental image undergoes the automatic updating of your frame of reference with your movement that is characteristic of moving in a real environment). These are among the reasons for saying that images are in an allocentric frame of reference and not in a cortico-centric frame of reference. What exactly this means and how it is possible for the spatial properties of represented space to be outside your head is the subject of this chapter.

Saying that the spatial properties of mental representations derive from properties of extra-personal space, rather than from properties in the brain does raise some problems. To get a feel for the sorts of problems it raises, imagine a simple shape, say a rectangle. Pick out some part of it, say the left vertical side. There is nothing problematic about being able to focus your attention on any part, say *that* part. Now think about some arbitrary place inside the rectangle, for example *there*. There is something odd about that. Although the imagined figure is out in the world, where is the *there* to which you just referred? Since the particular rectangle does not actually exist, where are the places in it located? One can think of a representation of a rectangle as a type of description – some set of proposition-like descriptors or some network, such as those used in artificial intelligence systems. That sort of a representation does contain parts that can be attentionally selected. But unlike parts of the rectangle, locations (especially empty locations) are not parts of a description nor is there any sense in which a description contains locations. The trouble is, as Gertrude Stein is said to have quipped about her childhood town of Oakland; There is no *there* there! Notwithstanding our phenomenology, there is no actual place that is picked out by the locative “there” when we refer to a place in our mental representation. Saying that it is a place in a representation is no help since we are trying to understand where the place you select with the locative *here* is represented as being. The place it is represented as being is not in some representation.

One obvious way to understand this apparent puzzle is to treat the representation as a mere fiction; Dorothy might have said that the Yellow Brick Road was *there*, in Kansas, even though Oz and everything in it are fictions. The problem with this option is that it does not easily explain why certain geometrical properties are true of the fictitious imagined figure. For example, when you slice the rectangle that you imagined earlier through one diagonal you can readily see that it yields two identical (congruent) right-angled triangles oriented 180 degrees from each other and located in two different places; one is *here* and the other *there* (the reader may want to check whether that is true in his or her imagination). If the rectangle were a mere fiction then why does slicing it create two fictitious triangles that are distinct but congruent tokens separated in the fictitious space by a rotation of 180 degrees? You might say that the two

parts created by the slice have the properties they have because in some real but implicit sense you *know* that this is what happens when you slice a rectangle. Exactly so, but still it is a unsatisfactory to say that the two triangles, which you know with some precision are now located *here* and *there*, do not have a location at all except in the way that the Yellow Brick Road has a location – especially since we have already agreed that the figures are *out there* in extra-personal space, and you can show that they remain *there*, fixed in allocentric space, as you move around them. In fact in this example, you can point to where the two triangles are located. It would be much more satisfying to actually attribute the *thereness* and the spatial relation between the two triangles to real space, even if the rectangle and two triangles are themselves fictions. It is their spatial coherence and their relative locations that is problematic, much more than the metaphysical existence of the figures themselves.

Two main options have been widely considered for how representations manage to exhibit the spatial properties discussed earlier. One option, which seems to have become the received wisdom in much of cognitive science and cognitive neuroscience, is that the spatial character (which includes the 6 properties listed in Section 5.4.1 above) derives from the spatial nature of the surface on which the representation is displayed, which many people think is on the surface of the first topographically organized region in the visual cortex – area V1. The other option is that it derives from the concurrently perceived world by a process that, with some hesitation, I have called *projection*. By “projection” I refer only to the hypothesis that certain real things in one’s perceptual field, which happen to be in roughly the same spatial configuration as the objects being imagined or recalled, are selected and indexed by FINSTs. These selected things are visually *marked*, as though there were visible labels assigned to them (an assumption that is inherent in most of the theories that have talked about visual marking, e.g., Ballard, Hayhoe, Pook, & Rao, 1997; Ullman, 1984; Watson & Humphreys, 1997; Yantis & Johnson, 1990; Yantis & Jones, 1991). What I *don’t* mean is that some sort of picture is projected onto the world, only FINSTs are projected, much as Plastic Man (in Figure 1-2) might stretch out his fingers and place them on things that are roughly in the same relative locations of things he is imagining, or as FINSTs might be attached to FINGs (in Figure 1-4 or Figure 2-2). This process occurs across many different modalities (including the proprioceptive modality, about which I will have more to say later). Below I sketch how such a mechanism allows the spatial properties of imagination to be inherited from spatial properties of the perceived world.

### 5.5.2 FINST Indexes can anchor spatial representations to real space

In earlier chapters I described the FINST index theory. It is perhaps illustrated most clearly in multiple-object tracking (MOT) studies discussed in chapter 2. In MOT, observers are able to select 4 or 5 briefly-cued objects (the “targets”), among a number of identical objects (the nontarget foils) and then to keep track of these targets as all objects move with unpredictable interpenetrating paths for some period of time, some 10 seconds or more. We have shown that these objects can be tracked without encoding their properties – indeed, changes in their color or shape are not noticed (Bahrami, 2003; Scholl, Pylyshyn, & Franconeri, 1999), and making them all different colors does not enhance tracking (Pylyshyn & Dennis, forthcoming). Targets can be tracked without significant decrement in performance even when they disappear by moving behind an opaque occluding surface or when they all disappear entirely for up to one second, as though you had blinked. We have shown that the Indexes (or FINSTs) can be assigned automatically by the onset of a new object, or voluntarily under certain conditions (when enough time is available to scan focal attention serially to each of the targets, see Pylyshyn & Annan, 2006, for details). We have also shown that indexes can be used to preselect a subset of items to search through (Burkell & Pylyshyn, 1997) and that this selection can withstand saccadic eye movement (Currie & Pylyshyn, 2003) or movement during multiple object tracking (Alvarez,

Arsenio, Horowitz, & Wolfe, 2005; Cohen & Pylyshyn, 2002). Selected items can also be enumerated very quickly, accurately and effortlessly, providing they can be automatically individuated (Trick, Audet, & Dales, 2003; Trick & Pylyshyn, 1994a). Although tracking seems like an attentionally-demanding task, we have reason to believe that there are certain other tasks known to require attention that can be carried out simultaneously with tracking without impairing tracking performance (e.g., monitoring for brief color changes or searching through targets for a specified letter, see Alvarez, Arsenio, Horowitz, & Wolfe, 2005; Leonard & Pylyshyn, 2003). Target-specific priming effects are also observed without the requirement to track (Haladjian & Pylyshyn, 2006; Noles, Scholl, & Mitroff, 2005). I have described some of these experiments in Chapter 1.

Given such a mechanism (illustrated in Figure 2-2), which allows stable access to a few individual things in the world and allows information relevant to each object to be associated with that object, we can bind the arguments of pattern-predicates and movement-commands to these objects (so that, for example, we can detect patterns such as collinearity among them or we can switch attention or gaze to them). This FINST mechanism provides a natural explanation for a number of phenomena that had led people to postulate an internal spatial medium or display. If we know (from instructions or from memory) the rough relative locations of a number of objects of thought or imagination, we can then associate these mental objects with indexed things in space. Once we are in such a relation to actual things in space, we are in a position to use our perceptual system to detect previously unnoticed patterns among these things or to scan our attention (or our gaze) from one to another indexed thing, or to judge relative distances between pairs of such things, and so on. Of course when vision is not available we must assume that the equivalent object-based perception in, say, the auditory or proprioceptive modalities takes its place. Since some of these non-visual modalities are importantly different from vision (e.g., their initial frame of reference is not the same as for vision) we will need to examine that option with some care. To introduce this topic, however, I begin with the visual case since it is not only easier to understand, it is also the most important modality for spatial orientation, as we will see.

Here is an example of an empirical phenomenon that can be accounted for in this way. One experiment mentioned briefly in the last chapter (and which has been frequently cited as providing a “window on the mind”, Denis & Kosslyn, 1999) is “mental scanning” – a phenomenon whereby the time to switch attention between imagined objects increases roughly linearly with the imagined distance between them. As I noted in Chapter 4, this phenomenon only appears when subjects take their task to require simulating, in whatever manner they can, a situation in which they are examining a real map, and the basic finding disappears if they take their task to be imagining that they viewing a map in which lights go on and off or when the attention scanning is downplayed, even in a context where they clearly need to switch their attention between two imagined objects (for details see Section 4.5.1). Still, under the appropriate conditions subjects do show this linear relation between distance and time so the question arises how they can simulate the observed reaction-time function (other than by generating a time interval without imagining any movement, which I don’t believe is what happens). Although this linear relation between distance and time has been interpreted as showing that mental images “have” metrical distances, it can also be seen in terms of scanning attention in real space between indexed objects. Here is how it might work.

If, for example, the map on the left of Figure 5-2 is imagined and a subject is asked to scan attention from one object (say the tree) to another object (say the tower), the time taken is proportional to the relative distance between the two imagined places.

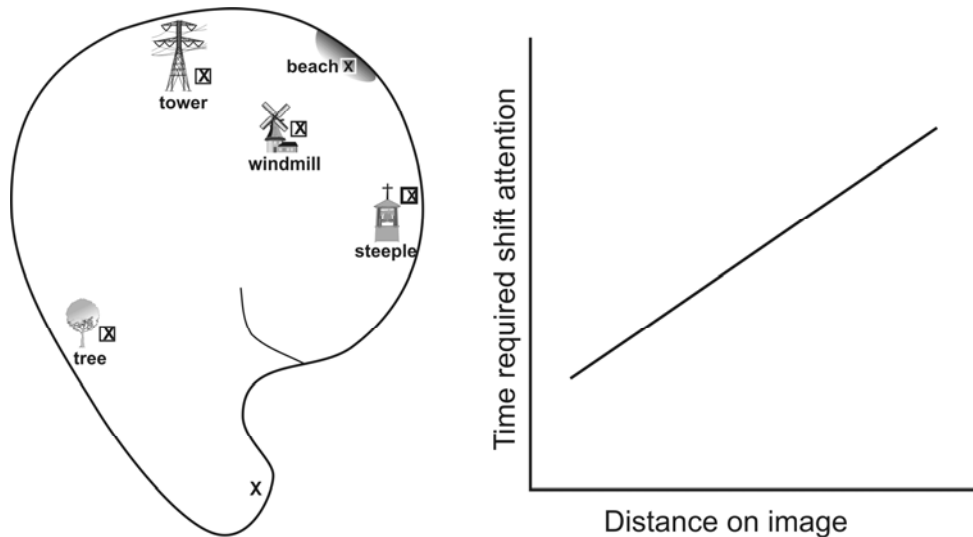


Figure 0-2: The task in this experiment is to learn the above map and then to imagine it while focusing on a specified location. Under appropriate conditions, the time it takes to shift attention to a second named place (and to “see” it in the “mind’s eye”) is given by the linear function shown on the right (Kosslyn, 1973).

But now suppose an observer is able to attach FINST Indexes to a few objects in the room where the experiment is being carried out, with the objects chosen so that their relative locations correspond roughly to memorized locations of objects on the map (we assume that one can memorize such locations in some unspecified qualitative configurational way that itself does not involve imagery). This sort of selection of concurrently perceived objects is illustrated in Figure 5-3. The subject would then be able to scan attention (or even direction of gaze) through physical space between the two relevant physical objects in the room. In that case the equation  $time = distance \div speed$  literally applies and the relevant time (give or take some fixed factor to account for the different psychophysical properties that might come into play when perceiving attentionally labeled objects). Thus this real scanning time would be proportional to the actually distance in space.

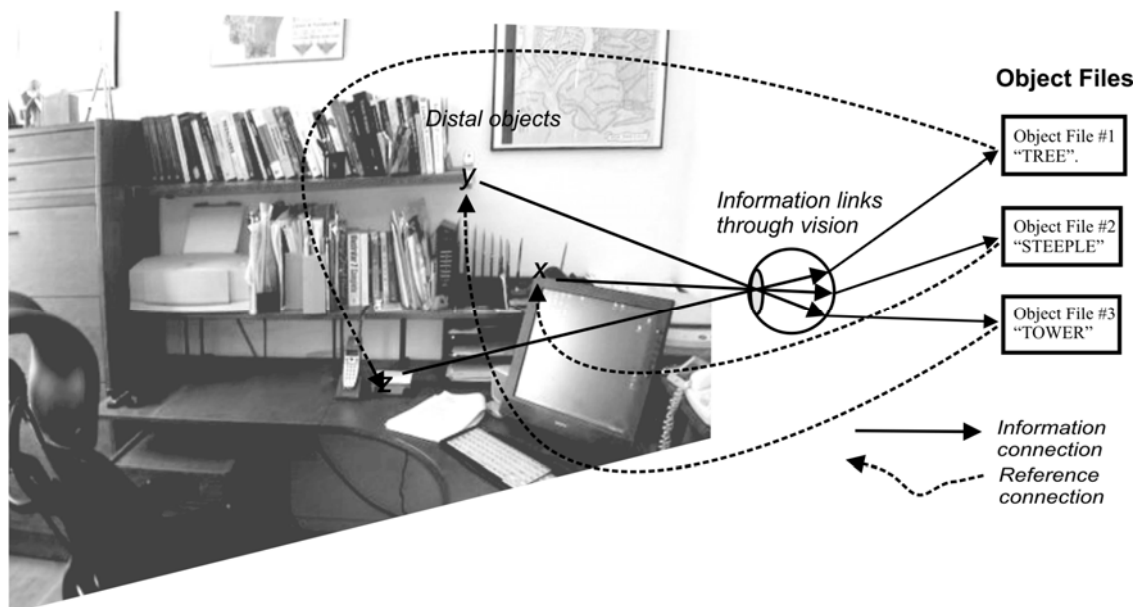


Figure 0-3. Binding indexes to objects in an office scene to associate these objects with the imagined mental objects (or, more accurately, indexed objects are associated with labels that indicate which objects are to be thought of as being at those locations). The indexing mechanism here is the same as illustrated in Figure 2-2.

The view I am proposing is that spatial properties of represented scenes derive from the spatial properties of selected (i.e., FINSTed) concurrently-perceived objects that are located in real space. The spatial properties I have in mind include the 6 properties listed above (in Section 5.4.1) as well as the many properties that have been cataloged in connection with studies of mental imagery or spatial representations in general (such as the example of mental scanning mentioned above, and also examples of problem solving that use spatial “mental models”, e.g., Johnson-Laird, 2001).<sup>60</sup>

This alternative story of how some mental representations get their the spatial properties is that they arise from the ability to select and keep track of individual objects in the world and then to visually detect their spatial patterns and also to perform visual operations on them, such as scanning focal attention between them. In chapters 1 and 2 I sketched a number of experiments illustrating nonconceptual selection and I described how FINSTs could provide an account of this process. These FINSTs may be usefully viewed as mechanisms of demonstrative identification. The assumption is that the small number of available FINST indexes can be captured by, or assigned to, salient things in the perceptual field, which then remain attached while either the individuals or the viewer moves. I suggested that this mechanism has widespread use in keeping track of a small selected subset of items of special interest (e.g., in order to search through them for a specified target), or to evaluate predicates for certain patterns (including such properties as collinearity and the cardinality of small sets of objects) and for maintaining the identity of individual objects across saccadic eye movements. These reference “pointers” can be assigned automatically by events in the visual world or voluntarily by serially scanning the items and thus selectively enabling the index grabbing operation to work on things that are selected by focal attention. Since indexes select individual objects, it is reasonable to assume that these selected objects can be distinguished by the visual system, just as the things one selects with focal attention can have a distinct visual role. Thus assuming that indexes can

<sup>60</sup> But for a critique of the interpretation of these mental model studies, see (Bonatti, 1994, 1998; Pylyshyn, 2003, Chapter 8).

provide input to the visual system, indexed objects can in effect be visually distinguished as though they were marked. This would allow the visual system to detect the spatial configuration among indexed objects. In this way indexes that make it possible to use the stable spatial properties of real objects located in real space to provide the spatial properties of represented space.

This idea also accounts for a number of findings cited in support of the picture theory of mental imagery. For example (Podgorny & Shepard, 1978) showed that there is a consistent pattern of time required to detect a small spot flashed at different places on a simple displayed pattern (for example it is faster when the spot occurs at the vertex or T-junction of a pattern such as the block letter F). They also found that the same reaction time pattern is observed when the display is no longer visible but merely recalled. Such a finding is compatible with the view that columns and rows that correspond to the projected pattern (e.g., the F) are simply indexed. The way that the visual system treats the imagined F pattern as being present during the imagery part of the experiment is by indexing the columns and rows that had contained the pattern. In general, findings that involve projecting an image into a visual scene are compatible with the assumption that certain geometrical features of the imagined pattern are effectively recreated by merely selecting features of the display where the image is supposed to be projected (including texture elements on the surface of the visual display that may not even be consciously noticed).

Another often-cited example is the the finding that perceptual-motor adaptation can be induced by using imagery rather than vision. In the classical adaptation paradigm, a subject wears prism goggles which shift the location of the perceived world by a fixed amount (say 23 degrees to the right). After attempting to reach for some objects on the table for just a few minutes the subject is able easily to correct for this displacement of the visual field. Moreover this correction becomes automatic so that when the goggles are removed the subject undershoots by the adapted amount (say by 23 degrees to the left). Ron Finke repeated this experiment without the subject seeing his or her arm (which was hidden under a surface) but with instructions to imagine that their hand was at the location where it would have appeared if they had worn the prism goggles. The same adaptation and undershoot was observed in this imagined displacement condition just as when prism goggles were worn. Although the experiment (described in, Finke, 1979) involved other controls, the one feature that it could not control for is the fact that the adaptation effect depends entirely on a mismatch between seen position of a hand and its felt position. All one needs in order to induce adaptation is some indication of the hand's location (which, as in the visual adaptation case, is different from the position where it was felt to be). No other visual property of the hand is required. Indeed, (Mather & Lackner, 1977) found that adaptation can be produced if subjects viewed a point source of light attached to their hand rather than the hand itself. In addition it appears that where the subject attends is important (Canon, 1970, 1971); even an immobile hand can produce adaptation providing the subject visually attends to it (Mather & Lackner, 1981). Thus the imagery condition in Finke's study provides all that is needed for adaptation, without any assumptions about the nature of imagery. In particular, subjects direct their gaze or their attention towards a particular (erroneous) location where they are in effect told to pretend their hand is located, thus focusing attention on the discordance between this viewed location and their kinesthetic and proprioceptive sense of the position of their arm. So all that is needed is a way to direct attention to a particular feature in the scene. So long as there is a visual feature at that location, to which a FINST can be attached, the results are easily explained.

Another visual-motor phenomenon that has attracted some interest from the pictorialist camp is the finding that a perceptual-motor compatibility effect (known as S-R compatibility or the Simon effect) can also be induced by visual imagery. It is well known that, regardless of what

the response signifies, responding in the direction of a relevant stimulus feature is faster than making a response away from the stimulus feature. For example, suppose that two stimuli are presented and a subject is asked to indicate whether one of them has a certain property (say the property of being the letter “A”) by pressing one of two buttons (one marked YES and the other marked NO), as shown in Figure 5-4. If the A is on the left and if the YES button is also on the left, the response is faster. If the YES response button had been on the right, or if the A target had been on the right the response would have been slower. Generally, responding towards the relevant visual pattern is faster. This effect is locationspecific rather than hand-specific since it holds even if the hands are crossed. What (Tlauka & McKenna, 1998) showed is that the same reaction time finding is obtained when the patterns (e.g., letters) are recalled after they are no longer visible or if subjects are instructed to imagine them at the compatible or incompatible locations. Thus the same pattern of spatial compatibility is observed with mental images as with actual visual perception. This too can be easily accounted for if one assumes that the locations where the letters are imagined is indexed (by indexing some visible feature at that location) and the letter name associated with the appropriate locations.

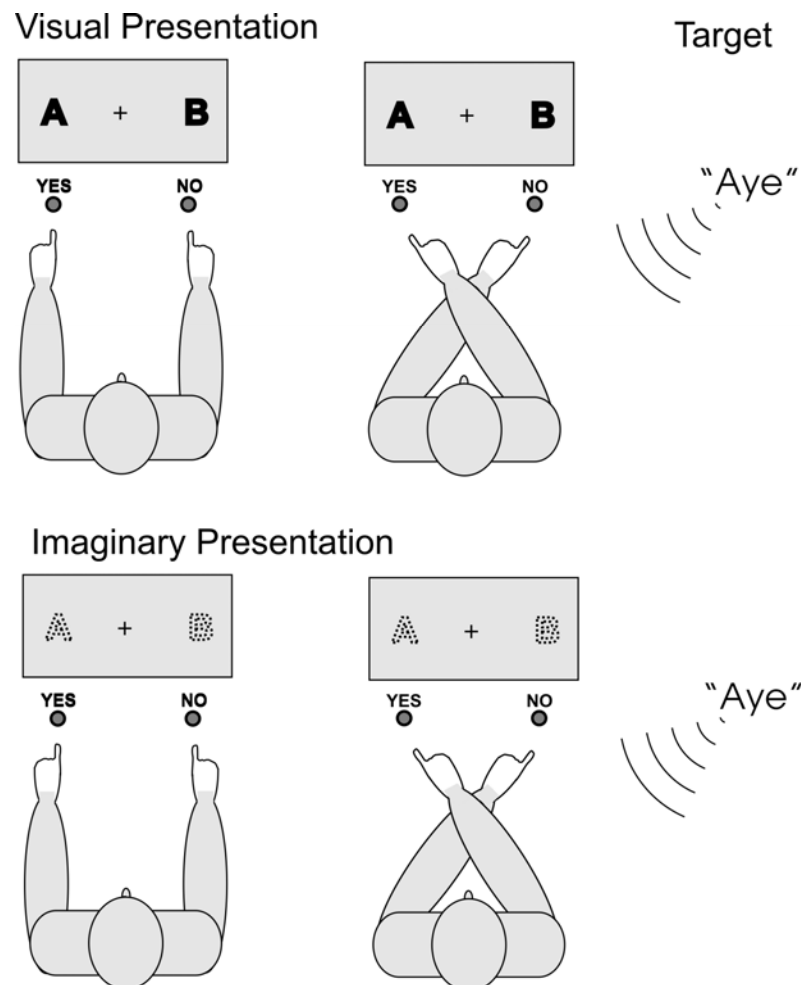


Figure 0-4. The S-R compatibility effect (this version is called the “Simon Effect”). Response (with either hand) towards location of the discriminating stimulus is faster than towards the irrelevant stimulus. This is true both for visual presentation (top 2) and for an imagined (or recalled) presentation. (This illustration is based loosely on Tlauka & McKenna, 1998, although it differs from their design which uses more complex stimuli).

In many experiments there are no obvious objects located at the relevant locations. Does this mean that, contrary to our earlier assumption, an empty location is indexed?. Not necessarily. There are two possibilities as to what can determine allocation of indexes in these cases. One is that there is always some feature near enough to the relevant places where the index is to be placed (keep in mind that the locations might be quite inaccurate since they might come from a qualitative recall). If you were to view digitized luminance-levels of a scene (such as show in computer vision texts like, Ballard, 1982) you would see discontinuities near almost every place in a scene – even in regions of blank walls. Since, by assumption, FINST allocation occurs early on in vision, it would have access to such pre-constancy discontinuities – brightness differences that have not been adjusted by whatever processes smooth out and replace the jagged input. Another possibility is that one can select empty locations in certain limited conditions, in particular we can pick them out relative to indexed objects in special cases. For example it may be that one can index an empty space if it can be specified as, say, being midway between two indexed objects. Although this is speculative there are no principled reasons for excluding the possibility of such special cases.

Clinical cases also suggest that representations constructed from memory work differently when they are accompanied by visual perception than when they are constructed in the dark or with eyes closed. Take, for example, the famous “duomo” experiments in which Edoardo Bisiach (Bisiach & Luzzatti, 1978) showed that patients with Unilateral Spatial Neglect (USN) tend not only to neglect one side of their visual world, but also the same side of their recollected or imagined world. What was particularly remarkable is that the patients could recall information from both sides of their visual field, as shown by the fact that if asked to imagine the same scene from a perspective rotated by 180 degrees, these patients accurately reported the formerly missing details. This was interpreted as showing that one side of an internal spatial display might be damaged in patients with USN. However in a recent interesting twist to this result (Chokron, Colliot, & Bartolomeo, 2004) showed that the neglect in mental images only occurs when the patients are simultaneously *viewing* a scene and does not appear when they imagine it with eyes closed. This supports the idea (championed by Bartolomeo & Chokron, 2002) that neglect is a deficit in attention control. There is also evidence that certain proprioceptive inputs can alter the axis of the imaginal neglect in USN patients, since neglect can be shifted with a shift in felt body orientation. (Karnath, Christ, & Hartje, 1993) showed that either turning the body by 15 degrees or decreasing the proprioceptive input by using vibrators applied to the neck muscles on the contralateral side, reduces the neglect. Also USN patients’ tactile exploration in the dark shows that the explored field is shifted in the direction away from the neglected side but the search pattern remains symmetrically distributed about the shifted direction and voluntary saccades do not show eye-centered directional deficits. Moreover, blocking proprioceptive signals over a series of treatments (using neck muscle vibrations) resulted in long-term reduction of USN symptoms (Schindler, Kerkhoff, Karnath, Keller, & Goldenberg, 2002). This and other results are consistent with the general thesis that the spatial representation (and its deficit in USN) is sensitive to concurrent spatial inputs (especially visual and proprioceptive).

These results also support the claim that in spatial mental imagery the spatial properties (including the distinction between left and right sides of the imagined scene) arise from the actual location in the currently perceived scene, of objects with which imagined objects are associated. There are many demonstrations in non-clinical patient populations showing that perception and recall of location is sensitive to concurrent spatial information. For example recall or orienting to auditory stimuli while visually viewing a scene improves performance compared with performance in the dark or without visual patterns (Warren, 1970). Spatial cues



based on ambient auditory stimulation with landmark auditory locations also improve report of the location of auditory stimuli (Dufour, Despres, & Pebayle, 2002). Conversely, viewing displays without any structure (as in the structureless display called a Ganzfeld, Avant, 1965) or being immersed in a impoverished sensory environment (e.g., an environment with low gravity, Watt, 1997) , results in spatial disorientation.

The data cited above provides further corroboration of the involvement of concurrently-perceived spatial information in the process of realizing what might be called the “spatial sense.” I should point out that although most of these examples involved concurrent visual patterns, we are able to represent spatial information (and demonstrate the 6 spatial conditions listed in section 5.4.1 above) without concurrent vision. Later I will discuss the question of how nonvisual information can serve the required anchoring function.

## 5.6 Projection in non-visual modalities

### 5.6.1 The unitary frame of reference assumption

As with other seductions of conscious contents, the fact that we have a unitary conscious spatial experience makes it tempting to assume that we have a single global representation of space in an allocentric frame of reference. There is also the independent assumption that this frame of reference is amodal since information from several modalities can be assimilated into a global frame of reference (or in some cases it is assumed that vision serves as the single frame of reference, since there is some evidence that auditory localization may be referred to a visual frame of reference, see Warren, 1970). However, there is also a great deal of evidence that we maintain a large number of different representations of spatial locations in different frames of reference (Colby & Goldberg, 1999). We know that the gaze-centered frame of reference plays an important role in visual-motor coordination (Crawford, Medendorp, & Marotta, 2004; Henriques, Klier, Smith, Lowy, & Crawford, 1998; Medendorp, Goltz, Villis, & Crawford, 2003; Snyder, Grieve, Brotchie, & Andersen, 1998) but even within this frame of reference the actual coordinates are modified extensively on-line by what are called gain fields (Salinas & Thier, 2000) which reflect head and body position, and even by not-yet-executed intentions to move the eyes (Andersen, 1995b; Duhamel, Colby, & Goldberg, 1992). There is also the famous distinction between ventral and dorsal visual systems which use different frames of reference, illustrated most famously by patients such as DF reported in (Milner & Goodale, 1995). These findings show that even within one modality, different functions (e.g., motor control and object recognition) involve different frames of reference – with the ventral system using a relatively local frame of reference (possibly object-centered for ease of object recognition) which represents qualitative rather than metrical spatial relations, as opposed to the dorsal system which uses many frames of reference, but primarily body centered frames of reference, and represents relatively precise spatial magnitudes (see, Bridgeman, Lewis, Heit, & Nagle, 1979). These representations (or what I called quasi-representations in Section 3.2.2) are likely laid out as neural layouts of activity. Note that the existence of neural layouts is not in question, only the frequent additional assumption that there is a layout that is global, allocentric and conscious – a layout that corresponds to the conscious experience of perceived extrapersonal space.

The use of multiple frames of reference is also illustrated by cases of unilateral spatial neglect (USN) – the deficit, discussed above, in which patients with damage in their parietal cortex fail to notice or respond to objects in half of their visual field. Even so clearly a spatial deficit appears to show the many different frames of reference that may be involved.

“Neglect occurs in all sensory modalities and can be expressed relative to any of several spatial reference frames, including retinocentric, body-centered, and environment-centered

can be specific for stimuli presented at particular distances. Some patients tend to ignore stimuli presented near the body, in peripersonal space, while responding normally to distant stimuli, or vice versa... Distance-specific neglect may be apparent only when the subject must produce a motor response to the stimulus, and not when spatial perception alone is tested” (Colby & Goldberg, 1999, p320-321).

Others have also found that more than one frame of reference may be affected in USN – in fact neglect may occur simultaneously in several frames of reference. For example (Behrmann & Tipper, 1999) reported that a patient showed neglect in both visual and object-centered frames of reference at the same time.

Properties of many of these frames of reference have been investigated, often with surprising results. For example, there appear to be integrated visual-tactile reference frames in peripersonal space, the near space around the body, especially surrounding the hand and face (Graziano & Gross, 1995; Làdavas, 2002). Visual stimuli presented near the body tend to be processed together with tactile stimuli at those locations, so that when one modality shows deficits, such as extinction (a deficit in processing two stimuli presented together bilaterally when neither is impaired when tested individually), the other tends to show similar deficits. The visual deficits in these cases are in a frame of reference relative to a body part (e.g., the hand or the face). The visual experience of the region around a body part appears to be tied to the somatosensory experience of the body part itself, so that it moves with the body part, appears with “phantom limb” experiences of amputees, and has even been shown to be extended with tool use (so that the experience of the space of one’s limb is extended to include a tool being used to explore space, see Làdavas, 2002). There is also evidence for hand-based visual representations or frames of reference<sup>61</sup> as well as gaze-centered and head-centered representations or frames. For example, pointing performance without vision is poorer when the starting position of the hand is not visible and the undershoot errors along the sagittal plane are larger when the hand begins from a point further from the body (which Chieffi, Allport, & Woodin, 1999, interpreted as suggesting that the visual target was encoded in hand coordinates) and, as noted above, auditory localization is poorer without spatial visual stimulation, such as provided by a textured surface, which many writers interpret as indicating that auditory information is encoded in a visual frame of reference (Warren, 1970).<sup>62</sup>

The visual and motor frames of reference are very closely linked. For example, the accuracy of pointing to a seen object after the eyes are closed is surprisingly high and persists for more than 15 seconds after the eyes are closed. More importantly, as mentioned earlier, the accuracy of pointing from a different imagined location is very poor *unless the subject actually moves to the new location*, even without vision during the move (Farrell & Thomson, 1998). It seems that many coordinate systems are automatically updated when we move. Of course these examples

---

<sup>61</sup> In the interest of expository simplicity I sometimes refer to neural layouts using the common convention of calling them representations, notwithstanding my earlier proposal that these may not be representations in the strong sense (see section 3.2.2(2)). A more neutral but more ponderous term might be “neural layout”.

<sup>62</sup> While this is a reasonable hypothesis, others (Dufour, Despres, & Pebayle, 2002) have subsequently shown that an auditory reference point also improves auditory localization, suggesting that it is not the visual frame of reference *per se* that is responsible for improved auditory localization, but the presence of perceptual anchors. This is consistent with the present thesis that localizations and other spatial skills rely on concurrent spatially structured perceptual stimulation.

of multiple frames of reference do not show that we don't also have a unitary frame of reference. However, if there is such a global frame of reference, it does not displace the individual ones and in fact would have to exist at the same time as the individual frameworks. Perhaps the main argument against the global frame of reference is that it is not needed given the prevalence of coordinate transformation functions described in the next section.

### 5.6.2 The role of coordinate transformations in externalizing space

One of the main motivations for the assumption that there is a uniform global frame of reference available to cognition is the fact that we easily go back and forth between perceptual modalities and, more importantly, between perception and motor action. Since perception begins with various peripheral frames of reference (e.g., vision starts with a retinal image, audition with an image of pitch on the basilar membrane) and motor control eventually requires a body-centered or world-centered frame of reference, an obvious solution is to convert everything into a common allocentric reference frame. Such a unitary coordinate system could then serve as the *lingua franca* for representing locations accessible to all modalities and effectors. This also comports well with our experience of seeing things in a stable allocentric frame of reference, as required by our unitary sense of first-person consciousness. But there are problems with this view. Motor control is not in an allocentric frame of reference. Commands must be issued in many different frames, including joint-angle frames and limb-centered frames (e.g., there is also evidence for coding in a hand-centered frame of reference). There are also many intermediate frames involved. For example in vision there are not only two retinal frames but also a cyclopean frame (which combines patterns from the two eyes into a 3D or stereo frame of reference) as well as a full 3D frame that is relative to a more distant (e.g., room) frame of reference.

There is also evidence that representations in different frames of reference are not deleted after conversion to another frame of reference occurs. Many of these intermediate representations leave their mark on ultimate motor performance, showing that intermediate frames of reference co-exist with downstream frames of reference (Baud-Bovy & Viviani, 1998). An interesting case occurs when we reach for an object after a brief exposure. While we are able to reach successfully even when eye movements occur between perceiving the object and reaching, both the retinal location and the motor representations are relevant in determining the outcome, as can be seen from errors in reaching that occur with intervening eye movements. Analysis of reaching errors reveals that "motor error commands cannot be treated independently of their frame of origin or the frame of their destined motor command" (Crawford, Medendorp, & Marotta, 2004, p10). It seems that the retinal location affects the direction of reach (see also, Batista, Buneo, Snyder, & Andersen, 1999). Many other studies confirm the residual effect of the multiple reference frames involved in the entire process, thus suggesting that a single conversion to a global frame of reference, if it occurs at all, cannot be the whole story.

An alternative to translating the many frames of reference into a uniform global frame is to carry out the translation pair-wise for the relevant frameworks. While this may sound like an inefficient way to do things, there is reason to think that special mechanisms are able to do this transformation efficiently. In fact a large number of coordinate-transformation sites have been identified through neurophysiological studies (many of these are described by Gallistel, 1999), consistent with coordinate transformations being a basic operation in the central nervous system (see also, Andersen, 1995a; Bizzi & Mussa-Ivaldi, 1995; Gallistel, 1999; Ghahramani, Wolpert, & Jordan, 1996; Snyder, 2000). Such transformations occur not only between modalities, but also between many distinct and constantly-changing frames of reference within a modality. Thus in moving your arm to grasp a perceived object you not only have to coordinate between visual

location information and proprioceptive location information, but also between a frame of reference in terms of joint angles to a frame of reference in terms of body-centered spatial coordinates and then from body-centered coordinates to allocentric coordinates. Since in reaching for something you generally move your eye, head and body (thus dynamically changing the body-centered coordinates), the coordination must occur rapidly on line. Although one might in principle convert each of these frames of reference to one global (e.g. allocentric) frame of reference, neurophysiological evidence appears to support pair-wise coordinate transformations among closely connected frameworks (e.g. eye-centered and head centered frames of reference to a body centered frame of reference or joint-angle frame of reference to a body-centered frame of reference).

There is evidence that the many frames of reference are tied together by a web of coordinate transformation operations. By contrast there is no known mechanism that operates directly in a global framework<sup>63</sup>, but rather there seem to be transformations only among cascades of frameworks beginning with receptor surfaces and ending with relevant effectors. Moreover, there are reasons to think that we do not need a single intermediary representation for coordinating spatial representations across modalities and between perception and action but rather that such coordination may operate by a series of coordinate transformations that are modulated by attention and intention. In their review of spatial representation in the brain, (Colby & Goldberg, 1999, p 319) suggest that a "... direct sensory-to-motor coordinate transformation obviates the need for a single representation of space in environmental coordinates."

Coordinate transformations are central in allowing perceptually-based frames of reference to coordinate with one another and to coordinate with action systems in the brain. In the motor system, coordinate transformations implicitly define equivalence classes of gestures from a given starting position to a fixed endpoint. These are just the classes  $\{S_i\}$  that Poincaré had postulated. Such equivalence classes accomplish for proprioception what perceived space does for vision – they provide unique proprioceptive information for unique places in extrapersonal space. As with vision, however, only a small subset of these – the occupied places – are indexed by a generalization of FINSTs to the proprioceptive modality (which in Pylyshyn, 1989, I referred to as *Anchors*)

The possible concern that such pair-wise transformations are an inefficient way to get from stimuli to responses is ameliorated by the suggestion that only a limited number of objects in different frames of reference are translated or transformed on line *as needed*. The plurality of reference frames, the speed with which we generally have to coordinate across such frames of reference make this idea of selective transformation plausible. A great deal of evidence, both behavioral and neurophysiological, suggests that only a small portion of perceptual information is selected and, moreover, very few objects need be converted to a motor frame of reference. The richness of the perceptual input and the complexity of the transformations that would be involved if the entire contents of each reference frame were converted also argues that this is not

---

<sup>63</sup>This is notwithstanding the remarkable work of (O'Keefe & Nadel, 1978) and others studying the hippocampus, who have demonstrated the existence of "place cells" in the rat hippocampus that respond selectively to places in an allocentric frame of reference. Place cells fire when the rat is at a particular place in a room (in the primate there are also "spatial view cells" that reportedly fire when the monkey is merely looking at the specific place, Rolls, 1999). These cells may also respond to a number of other properties (e.g., the animal's head direction), but nobody has suggested that they fire when the animal is planning a route. It is not known whether (or how) the hippocampus place cells *could* be used in constructing a plan which would, for example, need to represent the relative distance and direction of A and B when the animal is situated at some different place C. Some animals' ability to plan multi-destination routes is amazing (Gallistel & Cramer, 1996), but it is not clear how the hippocampus place cells could play a role in this computation. Without some idea how an animal could use the hippocampus representations to plan a route, calling that representation a "map" is misleading – it's more like a very simple Global Positioning System (GPS) that indicates when the animal is in the cell's "firing field" in a room but does not provide directions and distances between distant waypoints.

only computationally intractable, but it is unnecessary given the selective nature of the properties that go into the sensory-motor control process. This conclusion was also reached in a recent review by (Colby & Goldberg, 1999) who argued that attention plays an important roll in determining which objects are selected and converted. For similar reasons (Henriques, Klier, Smith, Lowy, & Crawford, 1998) proposed a “conversion-on-demand” principle in which only objects involved in a particular planned motor action are converted from retinal coordinates. A great deal of the location information is retained in eye-centered coordinates, as modified by eye, head, and body position, until it is needed (Batista, Buneo, Snyder, & Andersen, 1999; Klier, Wang, & Crawford, 2001; Snyder, 2000), and this may even be the case for auditory inputs (Stricanne, Andersen, & Mazzoni, 1996). The same may well be true of other pairs of reference frames, such as cyclopean, joint-angle, and so on (Colby & Goldberg, 1999), where in each case only selected objects are converted. No global allocentric coordinate system need be constructed since one needs nothing more than reference to a few objects as postulated by FINST theory.

### 5.6.3 Proprioceptive FINSTs? The projection hypothesis in nonvisual modalities

The foregoing sets the scene for a possible answer to the question: How can the projection hypothesis work in the absence of vision? A simple answer is that we can index objects without vision by simply using acoustical and other sense modalities whose distal causes are also laid out in space. There is no principled reason why the sort of picture illustrated in Figure 5-3 could not be realized if there were sound sources as well as sound reflecting objects functioning the way light does in vision. We know that people can use sound to find their way around. Certainly blind people are able to localize objects – not only sound sources, but sound reflecting objects (Ashmead, Davis, & Northington, 1995; Voss, Lassonde, Gougoux, Fortin, Guillemot et al., 2004). Although accuracy is not as good in audition as in vision, spatial-hearing of auditory landscapes may be sufficient to establish several auditory FINST anchors for purposes of the projection hypothesis.

More interesting are the various processes grouped under the category of *proprioception* and motor planning or preparatory motor actions (including kinesthetic, tactile, haptic and other somatosensory modalities, as well as preparatory motor programs). These are involved in spatial representation – for example it is known that the motor system is involved in thinking about spatial layouts. I have already mentioned experiments demonstrating that in some circumstances motor actions impair our ability to examine spatial layouts. Take, for example, the task of recalling the shape of a pattern by indicating what left or right turns one would have to make in traversing its outer contour. In one study, performance was measured for two types of responses, verbal (saying “Left” or “Right”) or motor (pointing to the correct terms in a column of symbols). Lee Brooks (Brooks, 1968) found that performance was impaired when a motor response was made. While Brooks interpreted that result as showing that mental images are spatial, what it really shows is that spatial gestures made concurrently with the examination of a spatial pattern impairs performance, as would be expected given the projection hypothesis as applied to the combination of visual and motor involvement in the pointing response. The work on Unilateral Spatial Neglect mentioned earlier shows that the representation of space (or more precisely, the exercise of attentional control) is also sensitive to availability of concurrent vision (Chokron, Colliot, & Bartolomeo, 2004) and to concurrent motor actions (Robertson, Nico, & Hood, 1995). There are also data showing that certain spatial operations on represented space (e.g., “mental rotation”) are affected by concomitant motor actions (Wexler, Kosslyn, & Berthoz, 1998).

Thus there are many demonstrations that engaging the visual system or the proprioceptive and motor system has an impact on active spatial representations (ASPARs). But we need more

than that. We need to show that the proprioceptive and motor systems can serve as the source of anchoring of represented space to actual perceived space. Also since people do not have to make overt movements in order to think about spatial properties we need to show that there is stable location information corresponding to the target of *potential* or *intended* motor actions as well as the right kind of potential proprioceptive information.

#### 5.6.4 Motor control, proprioception and intention define places in allocentric space

I begin by characterizing what has sometimes been called the spatial sense, a competence that is essential for proprioception to operate in allocentric coordinates, and therefore to provide the anchors assumed by the Projection Hypothesis when vision is not available. People are able to orient very well in space even if they are deprived of sight and sound, although the mechanisms that make this possible are not well known (apart from the fact that it involves the proprioceptive and kinesthetic, and to some extent the vestibular system). For the purposes of the projection hypothesis what is required is the capacity to identify particular places in extrapersonal space using proprioception, and to issue motor commands to move to places in extrapersonal space. These basic capacities, which were a major concern for Poincaré, depend heavily upon coordinate transformation operations. It is these transformations that compute the equivalence of proprioceptive information and/or motor actions with a common allocentric end point. The extra-personal locus is required because we are interested in how it is possible to bind FINST indexes (or *Anchors*) to stable places in extrapersonal space, using the proprioceptive modality, so they can provide stable references from which the spatial character of mental space can be derived perceptually (as suggested in the visual examples discussed earlier, and illustrated in Figure 5-3). If you are incrementally constructing a representation of a geometrical figure, each part of the representation must remain in place as other parts are added. And they must remain fixed despite movements of the eyes, head or body that occur during the process of indexing different perceived objects. It is the coordinate transformation operations that make it possible for proprioception to operate in a stable extra-personal frame of reference. For example, it's what makes it possible to point with your right hand index finger to places that had been occupied by the index finger of your left hand shortly before. This capacity was demonstrated by (Baud-Bovy & Viviani, 1998) who showed that although there were constant errors in pointing to proprioceptively-marked places, the pattern of errors was the same whether the initial "locating" or marking hand was the same or different from the pointing hand. A less technical way to put this is that our ability to sense or to reach fixed points in space, independent of the position of our eyes, head, hand or body, gives us potential access to points in allocentric space.<sup>64</sup> Of these potentially accessible places we only need to access a small number – the indexed subset.

There is a great deal of evidence that the preparation for make a gesture is a separate stage in undertaking an action. It is a stage that has been well studied, both in psychophysics (Hannus, Cornelissen, Lindemann, & Bekkering, 2005; Libet, Gleason, Wright, & Pearl, 1983; Musseler & Prinz, 1996) and in neuroscience (Andersen & Buneo, 2002; Buneo, Jarvis, Batista, & Andersen, 2002; Ghahramani, Wolpert, & Jordan, 1996; Gnadt, Bracewell, & Andersen, 1991; Henriques, Klier, Smith, Lowy, & Crawford, 1998; Karnath, 1997; Stricanne, Andersen, & Mazzoni, 1996). In psychophysics there are measures of the time it takes to prepare a response-sequence that is

---

<sup>64</sup> At a recent conference on "Frames of Reference in Spatial Cognition" held in Paris (Nov 17-19, 2005), Yves Rossetti presented a talk ("Questioning reference frames for sensori-motor transformations: driving the hand to eye-centred locations in space") in which he admonished people who assumed that motor control and proprioception occurred in an egocentric frame of reference and provided reasons why they are more correctly viewed as being in an extrapersonal frame of reference, a conclusion that agrees with the present thesis.

distinct from the time it takes to execute it.<sup>65</sup> One of the reasons that people postulated the existence of a “program of motor commands”, as distinct from movement itself, is the simple fact that a series of movements can be carried out without any kinesthetic feedback to keep track of where in the sequence the person is at any given time (Fleishman & Rich, 1963; Gandevia & Burke, 1992) (the most dramatic example of the ability to carry out action sequences without any proprioceptive information is provided by deafferented patients, such as the one described in, Cole, 1995, who have developed the ability to preplan a sequence and to monitor its execution visually or in some cases without any feedback). For that to be the case the sequence would have to be specified in advance and simply run off when needed. All so-called ballistic movements are of this sort, as are rapid serial sequences like playing familiar over-learned runs on the piano or sequences on a keyboard.

The idea that the intention to perform an action is carried out separately from the execution of the intention is now well-accepted in neuroscience. A classical study by Jean-René Duhamel (Duhamel, Colby, & Goldberg, 1992) showed that just before an eye movement, a cell in the part of parietal cortex of a monkey called the lateral intraparietal area (LIP), whose receptive field would about to include its signature feature after the saccade, becomes active (i.e., responded readily to a signal in that part of the visual field) before the saccade is actually completed. Thus the shift in the location of the receptive field of the neuron precedes the actual saccade. This suggests that this cell, which normally fires when the feature to which it is responsive occurs in its receptive field, actually anticipates the location that its receptive field will occupy after the eye movement is completed, so it becomes sensitized in advance! Many people have interpreted this to mean that the entire visual field is remapped (as was assumed in earlier theories which claimed that eye movements produced an “efference copy” that shifted retinotopic information, von Holst & Mittelstaedt, 1971/1950). But there is no need to make that assumption. In the cases studied by Duhamel and his colleagues, only certain of the salient features in the visual field need to be remapped<sup>66</sup> just as (Henriques, Klier, Smith, Lowy, & Crawford, 1998) claimed in their “conversion on demand” principle. It seems, then, that a sparse representation of the motor/proprioceptive scene is available containing the “intended” but yet-unexecuted move. There are now many examples of brain activity (usually in parietal cortex – particularly in area LIP) associated with the intention to act (e.g., Gnadt & Andersen, 1988; Shadlen & Newsome, 1996). Can this map of intentional preparation-for-action serve part of the function required by the projection hypothesis?

### 5.6.5 Summary of evidence in support of the projection hypothesis

I have reviewed a range of evidence relevant to the psychophysical capacity to pick out sensory objects (including intentional ones) in proprioceptive and motor space. This evidence

---

<sup>65</sup> The earliest theorizing about such a preparatory stage occurs in (Broadbent, 1958) and the demonstration of this stage by chronometric analysis occurs in (Sperling, 1967) where the author postulates a special buffer (the R-buffer) to hold information in a form that is intermediate between an “iconic memory” and an articulated verbal “rehearsal memory”. This buffer was necessitated in modeling the process of reading a display of briefly-presented letters in order to match the different speeds of information intake by the iconic and rehearsal buffers. By using a masking pattern Sperling was able to show that letters could be encoded in some form in a very short time (at a rate of about 10 ms per letter) but could not yet be recoded into subvocal speech (since that takes about 200 ms/letter). That leaves the need for some sort of storage which Sperling suggests has the form of a “program of motor instructions” for pronouncing the letters, which is subsequently executed during the “rehearsal” and “response” stage.

<sup>66</sup> Jean-René Duhamel has suggested to me in conversation that this implication may have arisen from the fact that the publication reporting this phenomenon (Duhamel, Colby, & Goldberg, 1992) showed a drawing of the monkey’s view, which may have suggested that *all* the features in that figure were remapped. This was not intended and may well not be true. The retina does indeed contain a different (shifted) image after the saccade, but there is no reason to believe that all receptive fields of neurons in LIP are remapped to accommodate this shift. In fact it was argued in a slightly different context by (Henriques, Klier, Smith, Lowy, & Crawford, 1998) that only a few task-relevant objects are remapped and the selection depends on attention (Colby & Goldberg, 1999).

together with the general finding that concurrent spatial stimulation appears to be necessary for our sense of space, provides support for the projection hypothesis even when no visual input is available. This discussion has ranged over a variety of topics so it is appropriate now that we pause to examine the inventory of findings that support the projection hypothesis:

*(1) Concurrent spatial stimulation may be necessary for our sense of space*

There is reason to believe that without sensory information from the world around us we would lose not only our orientation but also our sense of space. If a scene is free of visual features, visual-motor coordination is lost after only a short time. These so-called ganzfeld condition (in which subjects see only a uniform featureless expanse) results in disorientation and seriously impaired of form perception after only 90 seconds. The observer finds it difficult to locate a small object in the ganzfeld because “The observer not only does not know where he is looking, he also does not know where he has been looking previously” (Avant, 1965, p252). The same is true in total darkness where we get the so-called autokinetic illusion in which a stationary point of light appears to move and one’s orientation is disrupted (see Levy, 1972). There is also the recent interesting finding that orientation to sound is better when there are visual and/or auditory landmarks, and the amelioration of Unilateral Spatial Neglect with mental images when the patient describes the scene with eyes closed (Chokron, Colliot, & Bartolomeo, 2004).

*(2) Reasoning with eyes closed may still rely on visual persistence of objects in space*

There is evidence of short-term persistence of detailed visuospatial information, and there may also be persistence in other modalities (including audition and proprioception). Estimates of the duration of such storage varies from a few hundred milliseconds to a few minutes. One of the earliest reports of visual persistence was by George Sperling (Sperling, 1960) who used the method of partial report to show that information sufficient to report letters was available for about 250 ms after the disappearance of the information from a screen. More recently (Krekelberg, 2001) reported that position (though not color or shape) was available for 180 ms after the stimulus was extinguished in moving displays. There are various forms of visual memory that persist even longer and appear to be connected even more intimately with the motor system. For example, (Tyrrell, Rudolph, Eggers, & Leibowitz, 1993) have shown that visual information sufficient for guiding motor movement (placing a luminous dot at the right place in 3D) persists for more than 15 seconds. David Ingle has recently reported what he calls “central visual persistences” that last for up to 15 seconds (Ingle, 2005). These visual persistences are reminiscent of eidetic images but are not as detailed or long-lasting. They are, however, clearly situated in extrapersonal space and are closely connected with the motor system (e.g., it seems that the images can be moved manually by moving the hand). At the other extreme of duration, (Ishai & Sagi, 1995) found that accurate spatial information about flanking visual masks – capable of causing enhancement of detection thresholds – was available for up to 5 minutes. Such examples, most of which involve sensory persistence of locations in extrapersonal space, suggest that observations with eyes closed may not really be observations without visual input. None of the experiments that I am aware of, involving spatial images, required the observer to wait a few minutes in the dark (or with eyes closed) before being tested so long-term visual persistence may still be a factor in anchoring images to a visually perceived scene (as suggested in Figure 5-3).

*(3) Locations of targets of intended movement are represented even if no movement occurs*

There is evidence that the motor system activates locations in the visual cortex that are the targets of intended actions, even when no movement occurs. Recall the single-cell study by (Duhamel, Colby, & Goldberg, 1992) mentioned earlier, which shows that a cell with a receptive field that is about to cover a feature to which it is sensitive, begins to respond before the eye



movement is completed. It has also been shown that visual search is influenced by intended actions on the target objects even if no actual actions occur (Bekkering & Neggers, 2002). Such findings show that the locus of intended actions have measurable consequences even when not executed. There is converging evidence for the representation, in posterior parietal cortex, of the location of targets of intended actions, in an extrapersonal frame of reference (Andersen & Buneo, 2002; Snyder, Batista, & Andersen, 2000). This shows that coordinate transformation operations on visual inputs occurs even without the actual execution of motor actions. Thus imagining potential actions towards the recalled location of objects in space may create a frame of reference in the space of potential motor commands (or in the space of intended actions). Such a sparse representation, involving only the loci of intended actions, might also serve the function of anchoring the projection of individual objects, as required by the projection hypothesis. These findings are also consistent with the hypothesis that in making saccades, observers encode only the target of the saccade (this saccade-target theory is described in Currie, McConkie, & Carlson-Radvansky, 2000).

*(4) Acoustical and proprioceptive signals provide anchors the same way that vision does*

There are many relevant properties in common between vision, audition and proprioception which suggest that the earlier discussion about how objects can be indexed in vision (Section 5.5.2) apply equally to other modalities. Vision and audition both present a vast array of information about the sensory properties of the spatial layout before us. Even though only a very small subset of the potential information is encoded, the information is there as a possibility; it is what James Gibson referred to as the “ambient array”. There is evidence (surveyed above) that audition has some of the same object-locating capacities as does vision. I suggested that proprioception and intended motor actions also present something like an ambient sensory array because, thanks to coordinate transformation operations, they can represent a small number of individuals in what is effectively an allocentric frame of reference. If we think of the salient information that is selected (and indexed) from this proprioceptive array as “objects” then they also provide the capacity to anchor thoughts in stable allocentric space.

Given all these considerations it is not too far-fetched to assume that when you imagine a spatial layout in the dark, it is meaningful for you to think demonstrative and locative thoughts, such as *this* or *that*, as we do with vision, where the demonstratives pick out filled places in the proprioceptive landscape or the potential (intentional) motor landscape. The evidence I have quickly surveyed makes such a premise plausible. In fact it is quite intuitive to think of places in an imagined scene as being in the space in front of you, guided not by a retinotopic *internal* mental image but by externally sensed proprioceptive inputs or even signals for intended (but not executed) actions.

## Conclusions

The research and theoretical ideas described in this book takes as its fundamental premise that the mind is tuned to the world in certain ways, ways that no doubt come down to us through eons of evolution. Elsewhere I have discussed some of the ways this manifests itself in what David Marr and others have called Natural Constraints. The principle appears in many other domains where it shows up as the capacity to compute certain functions that would be logically impossible as described were it not for certain innate structures that allow a special sort of “approximation” to the ideal function. The approximation is more than a rough guess or heuristic: It is a function that *in our sort of world* will be very near the ideal function. The standard example in vision is the process of reversing the mapping from 3D distal world to 2D proximal images so as to permit the recovery of 3D shape. In that case not only does the visual

system use additional cues based on contour, shading, motion etc. to compute the inverse mapping from proximal to distal layout, but the process that computes this mapping is inherently constrained so that the unlimited number of logically-possible constructions are not available. The constructions that result are typically unique in our kind of world (in our ecological niche). A simple example is the interpretation of line drawings as 3D shapes (e.g., Figure 4-1). The label-consistency constraint I spoke of in Section 4.2 ensures that a unique labeling is given to the figure in cases where we actually have a unique percept. Similarly in the case of grammar, an (approximately) unique set of grammatical rules is induced even though the sparse input available for language learning is compatible with an unlimited number of grammars. What ensures the uniqueness of the induced grammar is the innate constraint on humanly possible grammars embodied in *Universal Grammar*. UG prevents the process of abduction of language structure from considering all but one of the logically possible candidate grammars.

In the present context the mechanism in question is one that helps to solve the concept grounding problem as well as the problem of capturing the numerical identity of enduring individual things. I do not claim that I have a solution to these problems, or even a serious candidate. What I present is a set of ideas, based on empirical research in my laboratory and elsewhere, that suggests how these problems might be approached, in the spirit of the Natural Constraints idea. The proposal is that there is in the early visual system, a primitive mechanism which accomplishes two tasks: it individuates things in the visual scene, and it provides direct reference to a small number of them. In this statement, “individuates” means that the visual system parses the visual world and segregates things in space and time so they can be treated as enduring individuals. This entails not only carrying out a figure-ground segregation (which is segregation in space), but also solving the correspondence problem (which is segregation in time). By a “direct reference” I mean essentially a demonstrative reference or an opaque pointer or index (which I have called a FINST) that allows epistemic access to a small number of the spatially and temporally segregated individuals without specifying any of their properties. It is crucial to this theory that neither of these functions involve appeal to concepts – that is, individuating is done prior to any properties being encoded as falling under concepts. I have been careful to point out that a property not being encoded is not the same as it not being causally involved in the process: There are specific properties that cause a FINST index to be assigned and that enable it to keep track of the indexed individuals – but those properties just are not encoded and a representation of those properties is not used in carrying out those functions.

I have devoted a significant part of this book to explicating the notion of FINST indexes and demonstrating the need for such a mechanism in accounting for many diverse empirical phenomena. I have argued that the idea comes naturally out of the well-accepted notion of focal attention, although it does represent an extension of this notion both in the assumption that there are several indexes (while there is typically only one locus of attention) and in the assumption that properties of indexed things are not involved in the indexing process: such properties need not be encoded in the course of individuation nor are they used to support the process of tracking. The major use of indexes is for binding the indexed individuals to the arguments of mental predicates or motor commands (including the command to move focal attention to a particular individual). Once attention (which I treat here primarily as the focus of visual processing) is allocated to indexed individuals, their properties can be encoded and stored in association with the individual (the properties are stored in what are called “object files”).

Because only predicates whose arguments are bound to individuals can be evaluated, it follows that only properties of indexed individuals can be conceptualized. This leads to a strong claim about conceptual and nonconceptual representations. The claim is that the only nonconceptual contents there are in perception are demonstratives or FINST indexes.

Nonconceptual content consists solely of the demonstrative reference to a small number of objects. There are no rich nonconceptual representations of the sort postulated by most philosophers who speak of nonconceptual representation.

But what about the major motivation given in philosophy for postulating nonconceptual representations, namely to account for the discrepancy between the rich fine-grained content of perceptual experience and the much more restricted content of conceptual representations? Several answers are offered to this conundrum. First, much of the content of conscious experience is problematic, if not irrelevant, to explanations of how vision works. There are many reasons to take that view. The main one is that the content of conscious experience is a fallible source of evidence, like any other form of observation, that can be overridden by other evidence – and very often must be overruled because it is subject to many forces, from bias effects to unconscious confabulation even in the absence of what ought to be clear experiences. If the claim that only properties of indexed individuals are represented (except for very short times during which the inertia of sensors may provide some richness of information) is correct, it raises the question of how to account for the large quantity of information that is needed by the encapsulated processes in vision, language, action planning and executing, and other domains. In my view it remains a real possibility that this sort of information may have to be treated as a subspecies of concepts – the sub-personal concepts, which have many of the characteristics of concepts (e.g., they appear to be symbolically encoded and take part in a form of unconscious reasoning, such as involved in parsing natural language or in early vision), yet may be different in kind from the usual *personal* concepts – for example we are not aware of them and they do not take part in general reasoning – i.e. they are not “inferentially promiscuous” (to use Steve Stich’s term, Stich, 1978).

To illustrate how the use of conscious contents as a privileged source of evidence is both tempting and misleading, I sketch a number of places in cognitive psychology where it has been a serious problem and then I focus on two closely related cases in more detail. These cases not only make the point that accepting conscious contents as necessarily corresponding to the contents of some mental representation is a mistake, but they also help to develop the connection between FINST indexes and nonconceptual content. The first case illustrates the way that conscious content has led us astray in developing theories of the representations underlying mental images used in reasoning. The second case, which is closely related but in many ways deeper, is the question of how we “represent” and use spatial properties. I have put “represent” in scare quotes because the proposal I make in the last chapter is that we do not need to represent spatial properties in reasoning, beyond the way we represent any domain in terms of concepts. Nonetheless, we do need to account for the very special spatial properties that appear when we actively use spatial relations in reasoning spatially, as we do when we think about problems in geometry, when we examine images for previously-unnoticed geometrical properties, and when we try to account for some remarkable parallels between visual imagery and visual perception.

The proposal I make in the last chapter is an externalist one. I propose that all we need to account for the apparently spatial properties of mentally represented space is a rough qualitative schema specifying spatial relations among a few way stations that we can use to pick out individuals in the concurrently perceived spatial surround using FINST indexes (or what, for non-visual modalities, I call *anchors*). With this capability we can do what is rather similar to “projecting” an image onto the perceived world with indexed objects being treated as labeled individuals. This then allows us to scan attention or gaze from one indexed individual to another, to judge the pattern that they form, estimate their relative distances, and so on. The point being that these judgments are all done by the visual system because the pattern that is being examined is now *out there* in the perceived scene. This idea, though simple, raises many problems, such as

how such judgments can be done without vision (e.g., in the dark) which in turn raises the question of how the proprioceptive system is able to make such judgments without actually reaching out and touching things in the world. All these complications reduce to empirical questions such as the frame of reference in which proprioception and motor control operate. The major idea at this junction turns out to be the mechanism, well known by neuroscience, for transforming representations in one frame of reference into another frame of reference. This function is known as the coordinate transformation operation and has been well recognized and studied by neuroscientists. I argue (based on a proposal by Henriques, Klier, Smith, Lowy, & Crawford, 1998, called “transformation on demand”) that only a few selected individuals need to be considered when superimposing memory representations and perceptual representations and these can be handled by coordinate transformation operations.

This brings the analysis full circle, to where FINSTs and various types of representation (especially nonconceptual representations) are seen to be interrelated and support the picture of a representation-governed system, such as I advocated in (Pylyshyn, 1984), augmented by direct causal links in the tradition of (Dretske, 1991; Dretske, 1981; Fodor, 1990), bring together a number of long-standing puzzles. Although it does not provide a definitive solution, it at least points a possible way to resolving a number of long standing problems in cognitive science.

## References

- Aglioti, S., DeSouza, J. F. X., & Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Current Biology*, 5(6), 679-685.
- Allen, R., McGeorge, P., Pearson, D., & Milne, A. B. (2004). Attention and expertise in multiple target tracking. *Appl. Cognit. Psychol.*, 18, 337-347.
- Alvarez, G. A., Arsenio, H. C., Horowitz, T. S., & Wolfe, J. M. (2005). Are multiple visual tracking and visual search mutually exclusive? *Journal of Experimental Psychology: Human Perception and Performance*, 31(4), 643-667.
- Alvarez, G. A., & Cavanagh, P. (2005). Independent attention resources for the left and right visual hemifields. *Psychological Science*, 16(8), 637-643.
- Alvarez, G. A., & Scholl, B. J. (2005). How Does Attention Select and Track Spatially Extended Objects? New Effects of Attentional Concentration and Amplification. *Journal of Experimental Psychology: General*, 134(4), 461-476.
- Andersen, R. A. (1995a). Coordinate transformations and motor planning in posterior parietal cortex, Gazzaniga, Michael S. (Ed). (1995). *The cognitive neurosciences*. (pp. 519-532). xiv, 1447pp.
- Andersen, R. A. (1995b). Encoding of intention and spatial location in the posterior parietal cortex. *Cerebral Cortex*, 5(5), 457-469.
- Andersen, R. A., & Buneo, C. A. (2002). Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience*, 25, 189-220.
- Ashmead, D. H., Davis, D. L., & Northington, A. (1995). Contribution of listeners' approaching motion to auditory distance perception. *Journal of Experimental Psychology: Human Perception & Performance*, 21(2), 239-256.
- Attneave, F. (1959). *Applications of information theory to psychology : a summary of basic concepts, methods, and results*. New York: Holt.
- Attneave, F., & Block, G. (1973). Apparent movement in tridimensional space. *Perception & Psychophysics*, 13(2), 301-307.
- Attneave, F., & Farrar, P. (1977). The visual world behind the head. *American Journal of Psychology*, 90(4), 549-563.
- Avant, L. L. (1965). Vision in the ganzfeld. *Psychological Bulletin*, 64, 246-258.
- Ayers, M. (1997). Is physical object a sortal concept? A reply to Xu. *Mind & Language*, 12(3-4), 393-405.
- Bachmann, T. (1989). Microgenesis as traced by the transient paired-forms paradigm. *Acta Psychologica*, 70, 3-17.
- Bahrami, B. (2003). Object property encoding and change blindness in multiple object tracking. *Visual Cognition*, 10(8), 949-963.
- Baillargeon, R., & DeVos, J. (1991). Object permanence in 3.5- and 4.5-month-old infants: further evidence. *Child Development*, 62, 1227-1246.
- Ballard, D. H., & Brown, C. M. (1982). *Computer vision*. New Jersey: Prentice-Hall.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4), 723-767.
- Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge, England: The University Press.
- Bartolomeo, P., & Chokron, S. (2002). Orienting of attention in left unilateral neglect. *Neuroscience and Biobehavioral Reviews*, 26(2), 217-234.
- Batista, A. P., Buneo, C. A., Snyder, L. H., & Andersen, R. A. (1999). Reach plans in eye-centered coordinates. *Science*, 285(5425), 257-260.
- Baud-Bovy, G., & Viviani, P. (1998). Pointing to kinesthetic targets in space. *Journal of Neuroscience*, 18(4), 1528-1545.
- Baylis, G. C. (1994). Visual attention and objects: Two-object cost with equal convexity. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 208-212.
- Baylis, G. C., & Driver, J. (1993). Visual attention and objects: Evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 451-470.
- Behrmann, M., & Tipper, S. (1999). Attention accesses multiple reference frames: Evidence from unilateral neglect. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 83-101.

- Bekkering, H., & Neggers, S. F. (2002). Visual search is modulated by action intentions. *Psychological Science*, 13(4), 370-374.
- Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.
- Best, C. T., Studdard-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering acoustic coherence in acoustic patterns. *Perception and Psychophysics*, 45(3), 237-250.
- Biederman, I. (1987). Recognition-by-components: A theory of human image interpretation. *Psychological Review*, 94, 115-148.
- Bisiach, E., & Luzzatti, C. (1978). Unilateral neglect of representational space. *Cortex*, 14(1), 129-133.
- Bizzi, E., & Mussa-Ivaldi, F. A. (1995). Toward a neurobiology of coordinate transformations, Gazzaniga, Michael S. (Ed). (1995). *The cognitive neurosciences*. (pp. 495-506). xiv, 1447pp.
- Black, A., & Pylyshyn, Z. W. (2004). Developmental Differences in Multiple Object Tracking. Paper presented at the Vision Sciences Society, Sarasota.
- Blaser, E., Pylyshyn, Z. W., & Domini, F. (1999). Measuring attention during 3D multielement tracking. *Investigative Ophthalmology and Visual Science*, 40(4), 552 (Abstract).
- Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature-space. *Nature*, 408(Nov 9), 196-199.
- Block, N. (1995). On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences*, 18, 227-247.
- Block, N. (2001). Is the content of experience the same as the content of thought? In D. Emmanuel (Ed.), *Language, Brain, and Cognitive Development: Essays in honor of Jacques Mehler*. Cambridge, MA: MIT Press/Bradford Books.
- Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, 9(2), 46-52.
- Block, N. (Ed.). (1981). *Imagery*. Cambridge, Mass.: MIT Press, a Bradford Book.
- Bonatti, L. (1994). Propositional reasoning by model? *Psychological Review*, 101(4), 725-733.
- Bonatti, L. (1998). What the mental logic-mental models controversy is not about: Braine, Martin D. S (Ed); O'Brien, David P (Ed). (1998). *Mental logic*. (pp. 435-445). vii, 481 pp. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.
- Bonatti, L., Frot, E., Zangl, R., & Mehler, J. (2002). The human first hypothesis: identification of conspecifics and individuation of objects in the young infant. *Cognitive Psychology*, 44, 388-426.
- Bornstein, R., & Pittman, T. (1992). *Perception without awareness*. New York, NY: Guilford Press.
- Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9(1), 27-38.
- Brentano, F. (1995 /1874). *Psychology from an empirical standpoint*. New York: Routledge.
- Bridgeman, B. (1992). Conscious vs. unconscious processes: The case of vision. *Theoretical Psychology*, 2, 73-88.
- Bridgeman, B., Lewis, S., Heit, G., & Nagle, M. (1979). Relation between cognitive and motor-oriented systems of visual position perception. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 692-700.
- Broadbent, D. E. (1958). *Perception and Communication*. London: Pergamon Press.
- Brooks, L. R. (1968). Spatial and verbal components of the act of recall. *Canadian Journal of Psychology*, 22(5), 349-368.
- Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, 64, 123-152.
- Buneo, C. A., Jarvis, M. R., Batista, A. P., & Andersen, R. A. (2002). Direct visuomotor transformations for reaching. *Nature*, 416(6881), 632-636.
- Burkell, J., & Pylyshyn, Z. W. (1997). Searching through subsets: A test of the visual indexing hypothesis. *Spatial Vision*, 11(2), 225-258.
- Calis, G. J., Sterenborg, J., & Maarse, F. (1984). Initial microgenetic steps in single-glance face recognition. *Acta Psychologica*, 55(3), 215-230.
- Campbell, J. (1993). The role of physical objects in spatial thinking. In N. Eilan & R. A. McCarthy & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology*. New York: Oxford University Press.
- Campbell, J. (2002). *Reference and Consciousness*. New York: Oxford University Press.
- Campbell, J. (2003). Reference as attention. *Philosophical Studies*, ?, 265-276.

- Canon, L. K. (1970). Intermodality inconsistency of input and directed attention as determinants of the nature of adaptation. *Journal of Experimental Psychology*, 84(1), 141-147.
- Canon, L. K. (1971). Directed attention and maladaptive "adaptation" to displacement of the visual field. *Journal of Experimental Psychology*, 88(3), 403-408.
- Carey, S. (forthcoming). *The Origin of Concepts*. Cambridge, MA: MIT Press.
- Carey, S., & Xu, F. (2001). Infants' knowledge of objects: Beyond object files and object tracking. *Cognition*, 80(1/2), 179-213.
- Cavanagh, P. (1992). Attention-based motion perception. *Science*, 257, 1563-1565.
- Cavanagh, P. (1999). Attention: Exporting vision to the mind. In C. Taddei-Ferretti & C. Musio (Eds.), *Neuronal Basis and Psychological Aspects of Consciousness* (pp. 129-143). Singapore: World Scientific.
- Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, 9(7), 349-354.
- Cavanagh, P., Labianca, A. T., & Thornton, I. M. (2001). Attention-based visual routines: Sprites. *Cognition*, 80(1-2), 47-60.
- Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition*, 23, 149-178.
- Chiang, W.-C., & Wynn, K. (2000). Infants' tracking of objects and collections. *Cognition*, 75, 1-27.
- Chieffi, S., Allport, D. A., & Woodin, M. (1999). Hand-centred coding of target location in visuo-spatial working memory. *Neuropsychologia*, 37, 495-502.
- Chokron, S., Colliot, P., & Bartolomeo, P. (2004). The role of vision on spatial representations. *Cortex*, 40, 281-290.
- Chomsky, N. (1957). Review of B. F. Skinner's *Verbal Behavior*. In J. A. Fodor & J. J. Katz (Eds.), *The Structure of Language*. Englewood Cliffs: Prentice-Hall.
- Churchland, P. M. (1988). Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor. *Philosophy of Science*, 55, 167-187.
- Clark, A. (2000). *A Theory of Sentience*. New York: Oxford University Press.
- Clark, A. (2004). Feature placing and proto-objects. *Philosophical Psychology*, 17(4), 443-469.
- Cohen, E. H., & Pylyshyn, Z. W. (2002). Searching through subsets of moving items. *Journal of Vision*, 2(7), 541a.
- Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22, 319-349.
- Cole, J. (1995). *Pride and a Daily Marathon*. Cambridge, MA: MIT Press.
- Collett, M., & Collett, T. S. (2000). How do insects use path integration for their navigation? *Biological Cybernetics*, 82, 245-259.
- Collett, T. S., & Collett, M. (2002). Memory use in insect visual navigation. *Nature*, 3, 542-552.
- Crawford, J. D., Medendorp, W. P., & Marotta, J. J. (2004). Spatial Transformations for Eye-Hand Coordination. *J Neurophysiol*, 92, 10-19.
- Culham, J. C., Brandt, S. A., Cavanagh, P., Kanwisher, N. G., Dale, A. M., & Tootell, R. B. H. (1998). Cortical fMRI activation produced by attentive tracking of moving targets. *J Neurophysiology*, 80(5), 2657-2670.
- Currie, C. B., McConkie, G. W., & Carlson-Radvansky, L. A. (2000). The role of the saccade target object in the perception of a visually stable world. *Perception and Psychophysics*, 62, 673-683.
- Currie, C. B., & Pylyshyn, Z. W. (2003). Maintenance of FINSTs across eye movements. Unpublished ms available at <http://rucss.rutgers.edu/~ccurrie/TitlePage.html>
- Dalla Barba, G., Rosenthal, V., & Visetti, Y.-V. (2002). The nature of mental imagery: How "null" is the null hypothesis? *Behavioral and Brain Sciences*, 25(2), 187-188.
- Danckert, J., & Maruff, P. (1997). Manipulating the disengage operation of covert visual spatial attention. *Perception & Psychophysics*, 59(4), Inc.
- Dawson, M. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychological Review*, 98, 569-603.
- Dawson, M., & Pylyshyn, Z. W. (1988). Natural constraints in apparent motion. In Z. W. Pylyshyn (Ed.), *Computational Processes in Human Vision: An interdisciplinary perspective* (pp. 99-120). Stamford, CT: Ablex Publishing.
- De Soto, C. B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality & Social Psychology*, 2(4), 513-521.

- Deford, J. K. (2003). *The horizontal effect: A new perceptual anisotropy*. Deford, John Kevin: U Louisville, US.
- Dehaene, S. (1997). *The number sense : how the mind creates mathematics*. New York: Oxford University Press.
- Denis, M., & Kosslyn, S. M. (1999). Scanning visual mental images: A window on the mind. *Cahiers de Psychologie Cognitive / Current Psychology of Cognition*, 18(4), 409-465.
- Dennett, D. C. (1978). *Toward a Cognitive Theory of Consciousness*. In C. W. Savage (Ed.), *Perception and Cognition: Issues in the Foundations of Psychology* (Vol. 9). Minneapolis: Univ. of Minnesota Press.
- Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown & Company.
- Dennis, J. L. M., & Pylyshyn, Z. W. (2002). Effect of object discriminability on multiple object tracking. Paper presented at the Vision Sciences 2002, Sarasota, FL.
- Devitt, M. (2006). *Ignorance of Language*. Oxford, UK: Oxford University Press.
- Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General*, 129(4), 481-507.
- Dretske, F. (1991). *Explaining Behavior*. Cambridge, MA: MIT Press.
- Dretske, F. (1993). Conscious experience. *Mind*, 102(406), 263-283.
- Dretske, F. (2006). What change blindness teaches about consciousness, Invited talk to the annual meeting of the society for the scientific study of consciousness. Oxford, June 2006.
- Dretske, F. I. (1981). *Knowledge and the Flow of Information*. Cambridge, Mass.: MIT Press, a Bradford Book.
- Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*, 2(7), 254-262.
- Driver, J., & Spence, C. (2004). Crossmodal spatial attention: evidence from human performance. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 179-220). New York, NY: Oxford University Press.
- Dufour, A., Despres, O., & Pebayle, T. (2002). Visual and auditory facilitation in auditory spatial localization. *Visual Cognition*, 9(6), 741-753.
- Duhamel, J.-R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255(5040), 90-92.
- Edelman, S. (2001). Neural spaces: A general framework for the understanding of cognition. *Behavioral & Brain Sciences*, 24(4), 664-665.
- Eilan, N., McCarthy, R. A., & Brewer, B. (Eds.). (1993). *Spatial representation: Problems in philosophy and psychology*. New York: Oxford University Press.
- Enns, J. T., & Franconeri, S. L. (2006). Environmental representations in multiple object tracking. Paper presented at the Visions Sciences 2006, Sarasota, FL.
- Enns, J. T., & Rensink, R. A. (1990). Sensitivity to three-dimensional orientation in visual search. *Psychological Science*, 1(5), 323-326.
- Epstein, W. (1982). Percept-percept couplings. *Perception*, 11, 75-83.
- Evans, G. (1982). *Varieties of Reference*. Oxford, UK: Oxford University Press.
- Farrell, M. J., & Thomson, J. A. (1998). Automatic spatial updating during locomotion without vision. *Quarterly J Experimental psychology A*, 51(3), 637-654.
- Farrell, M. J., & Thomson, J. A. (1999). On-line updating of spatial information during locomotion without vision. *Journal of Motor Behavior*, 31(1), 39-53.
- Finke, R. A. (1979). The Functional Equivalence of Mental Images and Errors of Movement. *Cognitive Psychology*, 11, 235-264.
- Finke, R. A., & Pinker, S. (1982). Spontaneous imagery scanning in mental extrapolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(2), 142-147.
- Fleishman, E. A., & Rich, S. (1963). Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. *Journal of Experimental Psychology*, 66(1), 6-11.
- Flombaum, J. I., Scholl, B. J., & Pylyshyn, Z. W. (2006). "Attentional high-beams" in tracking through occlusion. *Journal of Vision*, 6(6), 765a.
- Fodor, J. A. (1965). *Explanation in Psychology*. In M. Black (Ed.), *Philosophy in America*. Ithaca: Cornell Univ. Press.



- Fodor, J. A. (1968). The Appeal to Tacit Knowledge in Psychological Explanation. *Journal of Philosophy*, 65, 627-640.
- Fodor, J. A. (1980a). Methodological Solipsism Considered as a Research Strategy for Cognitive Psychology. *Behavioral and Brain Sciences*, 3(1), 63-73.
- Fodor, J. A. (1980b). Reply to Putnam. In M. Piattelli-Palmarini (Ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, Mass.: Harvard Univ. Press.
- Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass.: MIT Press, a Bradford Book.
- Fodor, J. A. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.
- Fodor, J. A. (2003). *Hume Variations*. Oxford: Clarendon Press.
- Fodor, J. A. (2004). *Revenge of the Given*. Unpublished manuscript.
- Fodor, J. A. (forthcoming). *The Language of Thought Revisited*. Cambridge, MA: MIT Press.
- Fodor, J. A., & Pylyshyn, Z. W. (1981). How Direct is Visual Perception? Some Reflections on Gibson's 'Ecological Approach'. *Cognition*, 9, 139-196.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3-71.
- Fougnie, D., & Marois, R. (2006). Distinct Capacity Limits for Attention and Working Memory. Evidence From Attentive Tracking and Visual Working Memory Paradigms. *Psychological Science*, 17(6), 526-534.
- Franconeri, S., Hollingworth, A., & Simons, D. J. (2005). Do new objects capture attention? *Psychological Science*.
- Franconeri, S. L., Pylyshyn, Z. W., & Scholl, B. J. (2006). Spatiotemporal cues for tracking objects through occlusion. *Journal of Vision*, 5(6), 1102a.
- Freeman, J. T. (1955). Set versus perceptual defense: A confirmation. *Journal of Abnormal & Social Psychology*, 51, 710-712.
- Frisby, J. P., & Clatworthy, J. L. (1975). Learning to see complex random-dot stereograms. *Perception*, 4(2), 173-178.
- Gallistel, C. R. (1990). *The Organization of Learning*. Cambridge, MA: MIT Press (A Bradford Book).
- Gallistel, C. R. (1999). Coordinate transformations in the genesis of directed action, Bly, Benjamin Martin (Ed); Rumelhart, David E. (Ed). (1999). *Cognitive science. Handbook of perception and cognition* (2nd ed.). (pp. 1-42). xvii, 391pp.
- Gallistel, C. R., & Cramer, A. E. (1996). Computations on metric maps in mammals: Getting oriented and choosing a multi-destination route. *Journal of Experimental Biology*, 199(1), 211-217.
- Gandevia, S. C., & Burke, D. (1992). Does the nervous system depend on kinesthetic information to control natural limb movements? *Behavioral and Brain Sciences*, 15(4), 614-632.
- Gazzaniga, M. S. (1995). Consciousness and the cerebral hemispheres. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1391-1400). Cambridge, MA: The MIT Press.
- Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(7), 1293-1326.
- Gentaz, E., & Hatwell, Y. (1998). The haptic oblique effect in the perception of rod orientation by blind adults. *Perception & Psychophysics*, 60(1), 157-167.
- Ghahramani, Z., Wolpert, D. M., & Jordan, M. I. (1996). Generalization to local remappings of the visuomotor coordinate transformation. *Journal of Neuroscience*, 16(21), 7085-7096.
- Gibson, J. J. (1979). *An Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
- Gilchrist, A. (1977). Perceived lightness depends on perceived spatial arrangement. *Science*, 195, 185-187.
- Gilden, D., Blake, R., & Hurst, G. (1995). Neural adaptation of imaginary visual motion. *Cognitive Psychology*, 28(1), 1-16.
- Gnadt, J. W., & Andersen, R. A. (1988). Memory related motor planning activity in posterior parietal cortex of macaque. *Exp. Brain Res.*, 70, 216-220.
- Gnadt, J. W., Bracewell, R. M., & Andersen, R. A. (1991). Sensorimotor transformation during eye movements to remembered visual targets. *Vision Research*, 31(4), 693-715.
- Gold, M. E. (1967). Language identification in the limit. *Information and Control*, 10, 447-474.
- Goldman, A. (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.
- Goodale, M., & Milner, D. (2004). *Sight Unseen*. New York, NY: Oxford University Press.

- Goodale, M. A., Jacobson, J. S., & Keillor, J. M. (1994). Differences in the visual control of pantomimed and natural grasping movements. *Neuropsychologia*, 32(10), 1159-1178.
- Goodale, M. A., Pelisson, D., & Prablanc, C. (1986). Large adjustments in visually guided reaching do not depend on vision of the hand or perception of target displacement. *Nature*, 320, 748-750.
- Goodman, N. (1968). *Languages of Art*. Indianapolis: Bobbs-Merrill.
- Graziano, M. S. A., & Gross, C. G. (1995). The representation of extrapersonal space: a possible role for bimodal, visuo-tactile neurons. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (pp. 1021-1034). Cambridge, MA: MIT Press.
- Gul, E., & Atherton, D. (1989). Transputer implementation for multiple target tracking. *Microprocessors Microsystems*, 13(3), 188-194.
- Gumperz, J., & Levinson, S. (Eds.). (1996). *Rethinking Linguistic Relativity: Studies in the social and cultural foundations of language*. Cambridge, UK: Cambridge University Press.
- Gunther, Y. H. (Ed.). (2003). *Essays on Nonconceptual Content*. Cambridge, MA: MIT Press (Bradford Book).
- Hagenzieker, M. P., van der Heijden, A. H. C., & Hagenaar, R. (1990). Time course in visual-information processing: some empirical evidence for inhibition. *Psychological Research*, 52, 13-21.
- Haladjian, H. H., & Pylyshyn, Z. W. (2006). Implicit multiple object tracking without an explicit tracking task. *Journal of Vision*, 6(6)(6), 773a.
- Halsey, R., & Chapanis, A. (1951). On the number of absolutely identifiable spectral hues. *Journal of the Optical Society of America*, 41, 1057-1058.
- Hannus, A., Cornelissen, F. W., Lindemann, O., & Bekkering, H. (2005). Selection-for-action in visual search. *Acta Psychologica*, 118(1-2), 171-191.
- Hansen, B. C., & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision*, 4(12), 1044-1060.
- Haugeland, J. (1978). The Nature and Plausibility of Cognitivism. *Behavioral and Brain Sciences*, 2, 215-260.
- Heeley, D. W., Buchanan-Smith, H. M., Cromwell, J. A., & Wright, J. S. (1997). The oblique effect in orientation acuity. *Vision Research*, 37(2), 235-242.
- Henriques, D. Y. P., Klier, E. M., Smith, M. A., Lowy, D., & Crawford, J. D. (1998). Gaze-centered remapping of remembered visual space in an open-loop pointing task. *Journal of Neuroscience*, 18(4), 1583-1594.
- Hermer, L., & Spelke, E. (1996). Modularity and development: the case of spatial reorientation. *Cognition*, 61(3), 195-232.
- Hinton, C. H. (1906). *The Fourth Dimension*. London: George Allen and Unwin.
- Hirsch, E. (1997). Basic objects: A reply to Xu. *Mind & Language*, 12(3-4), 406-412.
- Hirstein, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. Cambridge, MA: MIT Press.
- Hochberg, J. (1968). In the mind's eye. In R. N. Haber (Ed.), *Contemporary theory and research in visual perception* (pp. 309-331). New York: Holt, Rinehart & Winston.
- Hoffman, D. D. (1998). *Visual Intelligence: How We Create What We See*. New York: W.W. Norton.
- Horn, B. K. P. (1986). *Robot vision*. Cambridge, MA and London: MIT Press.
- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195, 215-243.
- Huntley-Fenner, G., Carey, S., & Salimando, A. (2002). Objects are individuals by stuff doesn't count: Perceived rigidity and cohesiveness influence infants' representation of small number of discrete entities. *Cognition*, 85(3), 203-221.
- Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review*, 75(6), 550-560.
- Ingle, D. (2005). Central visual persistences: I. Visual and kinesthetic interactions. *Perception*, 34(9), 1135-1151.
- Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of attention. *Cognitive Psychology*, 4(3), 171-216.
- Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 18(2), 307-317.

- Irwin, D. E., & Gordon, R. D. (1998). Eye movements, attention and trans-saccadic memory. *Visual Cognition*, 5(1-2), 127-155.
- Ishai, A., & Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science*, 268(5218), 1772-1774.
- Jackson, F. (1997). *Perception: A representative theory*. Cambridge, UK: Cambridge University Press.
- Johnson, S. P. (2001). Visual development in human infants: Binding features, surfaces, and objects. *Visual Cognition*, 8(3-5), 565-578.
- Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences*, 5(10), 434-442.
- Jolicoeur, P., & Kosslyn, S. M. (1985). Is time to scan visual images due to demand characteristics? *Memory & Cognition*, 13, 320-332.
- Jovicich, J., Peters, R., Koch, C., Braun, J., Chang, L., & Ernst, T. (2001). Brain areas specific for attentional load in a motion-tracking task. *Journal of Cognitive Neuroscience*, 13, 1048-1058.
- Julesz, B. (1971). *Foundations of Cyclopean Perception*. Chicago: Univ. of Chicago Press.
- Kabanza, F., Barbeau, M., & St-Denis, R. (1997). Planning control rules for reactive agents. *Artificial Intelligence*, 95, 67-113.
- Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2), 175-219.
- Kanizsa, G., & Gerbino, W. (1982). Amodal completion: Seeing or thinking? In B. Beck (Ed.), *Organization and Representation in Perception* (pp. 167-190). Hillsdale, NJ: Erlbaum.
- Kappers, A. M. L., & Koenderink, J. J. (1999). Haptic perception of spatial relations. *Perception*, 28(6), 781-795.
- Karnath, H. O. (1997). Spatial orientation and the representation of space with parietal lobe lesions. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, 352(1360), 1411-1419.
- Karnath, H. O., Christ, K., & Hartje, W. (1993). Decrease of contralateral neglect by neck muscle vibration and spatial orientation of trunk midline. *Brain*, 116(2), 383-396.
- Kazanovich, Y., & Borisyuk, R. (2006). An oscillatory neural model of multiple object tracking. *Neural Computation*, 18(6), 1413-1440.
- Keane, B. (2004). *A Case for Visual Objects in Early Vision: A Response to Austen Clark's Theory of Sentience*. Unpublished PhD, Rutgers University, New Brunswick, NJ.
- Keane, B. P., & Pylyshyn, Z. W. (2006). Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. *Cognitive Psychology*, 52(4), 346-368.
- Klein, R. M. (1988). Inhibitory tagging system facilitates visual search. *Nature*, 334(6181), 430-431.
- Klier, E. M., Wang, H., & Crawford, J. D. (2001). The superior colliculus encodes gaze commands in retinal coordinates. *Nature Neuroscience*, 4(6), 627-632.
- Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4, 219-227.
- Koenderink, J. J. (1990). *Solid Shape*. Cambridge, MA: MIT Press.
- Kohler, W. (1947). *Gestalt Psychology, An Introduction to New Concepts in Modern Psychology*. New York: Liveright.
- Kolers, P. A., & von Grunau, M. (1976). Shape and color in apparent motion. *Vision Research*, 16(4), 329-335.
- Kosslyn, S. M. (1973). Scanning Visual Images: Some Structural Implications. *Perception and Psychophysics*, 14, 90-94.
- Kosslyn, S. M. (1975). The Information Represented in Visual Images. *Cognitive Psychology*, 7, 341-370.
- Kosslyn, S. M. (1978). Measuring the visual angle of the mind's eye. *Cognitive Psychology*, 10, 356-389.
- Kosslyn, S. M. (1980). *Image and Mind*. Cambridge, Mass.: Harvard Univ. Press.
- Kosslyn, S. M. (1981). The Medium and the Message in Mental Imagery: A Theory. *Psychological Review*, 88, 46-66.
- Kosslyn, S. M. (1994). *Image and Brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.
- Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 46-60.

- Kosslyn, S. M., Sukel, K. E., & Bly, B. M. (1999). Squinting with the mind's eye: Effects of stimulus resolution on imaginal and perceptual comparisons. *Memory & Cognition*, 27(2), 276-287.
- Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. New York, NY: Oxford University Press.
- Krekelberg, B. (2001). The persistence of position. *Vision Research*, 41(4), 529-539.
- Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.
- Krojgaard, P. (2004). A review of object individuation in infancy. *British Journal of Developmental Psychology*, 22(2), 159-183.
- Kuhn, T. (1957). *The Copernican Revolution. Planetary Astronomy in the Development of Western Thought*. Cambridge, MA: MIT Press.
- Làdavas, E. (2002). Functional and dynamic properties of visual peripersonal space. *Trends in Cognitive Sciences*, 6(1), 17-22.
- Leonard, C., & Pylyshyn, Z. W. (2003). Measuring the attentional demand of multiple object tracking (MOT). *Journal of Vision*, 3(9), 582a.
- Lepore, E., & Ludwig, K. (2000). The semantics and pragmatics of complex demonstratives. *Mind*, 109, 199-240.
- Leslie, A. M. (1982). The perception of causality in infants. *Perception*, 11, 173-186.
- Leslie, A. M. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception*, 13(3), 287-305.
- Leslie, A. M. (1988). The necessity of illusion: Perception and thought in infancy. In L. Weiskrantz (Ed.), *Thought Without Language*. Oxford: Oxford Science Publications.
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25, 265-288.
- Levine, J. (forthcoming). Demonstrative thought.
- Levy, J. (1972). Autokinetic illusion: A systematic review of theories, measures and independent variables. *Psychological Bulletin*, 78, 457-474.
- Lewis, D. (1971). Analog and Digital. *Nous*, 5(3), 321-327.
- Libet, B. (2004). *Mind time: the temporal factor in consciousness*. Cambridge, MA: Harvard University Press.
- Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*, 106(Pt 3), 623-642.
- Lindberg, D. C. (1976). *Theories of vision from al-Kindi to Kepler*. Chicago: University of Chicago Press.
- Liu, G., Austen, E. L., Booth, K. S., Fisher, B. D., Argue, R., Rempel, M. I., & Enns, J. T. (2005). Multiple-Object Tracking Is Based on Scene, Not Retinal, Coordinates. *Journal of Experimental Psychology: Human Perception and Performance*, 31(2), 235-247.
- Lormand, E. (1996). Nonphenomenal consciousness. *Noûs*, 30(2), 242-261.
- Luce, R. D., D'Zmura, M., Hoffman, D. D., Iverson, G. J., & Romney, A. K. (Eds.). (1995). *Geometric representations of perceptual phenomena: Papers in honor of Tarow Indow on his 70th birthday (Vol. 356)*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.
- Mack, A., & Rock, I. (1998). *Inattention blindness*. Cambridge, MA: MIT Press.
- Marr, D. (1977). Artificial Intelligence - A Personal View. *Artificial Intelligence*, 9, 37-48.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.
- Mather, J. A., & Lackner, J. R. (1977). Adaptation to visual rearrangement: Role of sensory discordance. *Quarterly Journal of Experimental Psychology*, 29(2), 237-244.
- Mather, J. A., & Lackner, J. R. (1981). Adaptation to visual displacement: Contribution of proprioceptive, visual, and attentional factors. *Perception*, 10(4), 367-374.
- McCloskey, M., & Kargon, R. (1988). The meaning and use of historical models in the study of intuitive physics: Strauss, Sidney (Ed). (1988). *Ontogeny, phylogeny, and historical development*. (pp. 49-67). xxi, 158 pp. Westport, CT, US: Ablex Publishing.
- McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4), 636-649.
- McDaniel, K. D., & McDaniel, L. D. (1991). Anton's syndrome in a patient with posttraumatic optic neuropathy and bifrontal contusions. *Archives of Neurology*, 48(1), 101-105.

- McMahon, M. J., & MacLeod, D. I. A. (2003). The origin of the oblique effect examined with pattern adaptation and masking. *Journal of Vision*, 3(3), 230-239.
- Medendorp, W. P., Goltz, H. C., Villis, T., & Crawford, J. D. (2003). Gaze-Centered Updating of Visual Space in Human Parietal Cortex. *The Journal of Neuroscience*, 23(15), 6209-6214.
- Meltzoff, A. N. (1993). Molyneux's babies: Cross-modal perception, imitation and the mind of the preverbal infant. In N. Eilean & R. McCarthy & B. Brewer (Eds.), *Spatial Representation: Problems in Philosophy and Psychology* (pp. 219-235). New York, NY: Oxford University Press.
- Menzel, R., Greggers, U., Smith, A., Berger, S., Brandt, R., Brunke, S., Bundrock, G., Hulse, S., Plumpe, T., Schaupp, F., Schuttler, E., Stach, S., Stindt, J., Stollhoff, N., & Watzl, S. (2005). Honey bees navigate according to a map-like spatial memory. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8), 3040-3045.
- Metcalfe, J., & Shimamura, A. P. (Eds.). (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.
- Michotte, A. (1946/1963). *The perception of causality*. Andover: Methuen.
- Milner, A. D., & Goodale, M. A. (1995). *The Visual Brain in Action*. New York: Oxford University Press.
- Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2006). Attentive tracking of multiple objects modulates neuronal responses in area V4 of the macaque. *Journal of Vision*, 6(6), 772-772.
- Moore, C., Yantis, S., & Vaughan, B. (1998). Object-based visual selection: Evidence from perceptual completion. *Psychological Science*, 9, 104-110.
- Mounts, J. R. W. (2000). Attentional capture by abrupt onsets and feature singletons produces inhibitory surrounds. *Perception & Psychophysics*, 62(7), 1485-1493.
- Mueller, H. J., & Rabbitt, P. M. (1989). Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption. *Journal of Experimental Psychology: Human Perception & Performance*, 15(2), 315-330.
- Musseler, J., & Prinz, W. (1996). Action planning during the presentation of stimulus sequences: effects of compatible and incompatible stimuli. *Psychological Research*, 59(1), 48-63.
- Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall.
- Nicod, J. (1970). *Geometry and Induction*. Berkeley: Univ. of California Press.
- Nisbett, R. E., & Valins, S. (1987). Perceiving the causes of one's own behavior. In E. E. Jones & D. E. Kanouse & H. H. Kelley & R. E. Nisbett & S. Valins (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 63-78). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259.
- Nissen, M. J. (1985). Accessing features and objects: Is location special? In M. I. Posner & O. S. Marin (Eds.), *Attention and performance XI* (pp. 205-219). Hillsdale, NJ: Lawrence Erlbaum.
- Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.
- Noles, N. S., Scholl, B. J., & Mitroff, S. R. (2005). The persistence of object file representations. *Perception & Psychophysics*, 67(2), 324-334.
- Oberle, C. D., McBeath, M. K., Madigan, S. C., & Sugar, T. G. (2005). The Galileo Bias: A Naive Conceptual Belief That Influences People's Perceptions and Performance in a Ball-Dropping Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(4), 643-653.
- Ogawa, H., Takeda, Y., & Yagi, A. (2002). Inhibitory tagging on randomly moving objects. *Psychological Science*, 13(2), 125-129.
- Ogawa, H., & Yagi, A. (2002). The effect of information of untracked objects on multiple object tracking. *Japanese Journal of Psychonomic Science*, 22(1), 49-50.
- O'Hearn, K., Landau, B., & Hoffman, J. E. (2005). Multiple Object Tracking in People with Williams Syndrome and in Normally Developing Children. *Psychological Science*, 16(11), 905-912.
- O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a cognitive map*. Oxford: Oxford University Press.
- O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, 46, 461-488.
- O'Regan, J. K., & Noë, A. (2002). A sensorymotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5), 939-1031.
- Paillard, J. (Ed.). (1991). *Brain and Space*. New York: Oxford University Press.
- Pashler, H. E. (1998). *The Psychology of Attention*. Cambridge, MA: MIT Press (A Bradford Book).

- Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press/Bradford Books.
- Peacocke, C. (1993). Intuitive mechanics, psychological reality and the idea of a material object. In N. Eilan & R. A. McCarthy & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology*. New York: Oxford University Press.
- Perry, J. (1979). The problem of the essential indexical. *Noûs*, 13, 3-21.
- Philipona, D., O'Regan, K., & Nadal, J.-P. (2003). Is there something out there ? Inferring space from sensorimotor dependencies. *Neural Computation*, 15(9).
- Podgorny, P., & Shepard, R. N. (1978). Functional representations common to visual perception and imagination. *Journal of Experimental Psychology: Human Perception and Performance*, 4(1), 21-35.
- Poggio, T., Torre, V., & Koch, C. (1990). Computational vision and regularization theory. In S. Ullman & W. Richards (Eds.), *Image Understanding 1989* (Vol. 3, pp. 1-18). Norwood, N.J.: Ablex Publishing.
- Poincaré, H. (1963/1913). Why space has three dimensions (J. W. Bolduc, Trans.), *Mathematics and Science: Last Essays* (pp. 25-44). New York: Dover.
- Posner, M. I. (1980). Orienting of Attention. *Quarterly Journal of Experimental Psychology*, 32, 3-25.
- Pylyshyn, Z. W. (1973a). The Role of Competence Theories in Cognitive Psychology. *Journal of Psycholinguistics Research*, 2, 21-50.
- Pylyshyn, Z. W. (1973b). What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin*, 80, 1-24.
- Pylyshyn, Z. W. (1979). The Rate of 'Mental Rotation' of Images: A Test of a Holistic Analogue Hypothesis. *Memory and Cognition*, 7, 19-28.
- Pylyshyn, Z. W. (1980). Cognition and computation: Issues in the foundation of cognitive science. *Behavioral and Brain Sciences*, 3(1), 111-132.
- Pylyshyn, Z. W. (1981). The imagery debate: Analogue media versus tacit knowledge. *Psychological Review*, 88, 16-45.
- Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press (Also available through CogNet).
- Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, 32, 65-97.
- Pylyshyn, Z. W. (1991). Rules and Representation: Chomsky and representational realism. In A. Kashir (Ed.), *The Chomskian Turn*. Oxford: Basil Blackwell Limited.
- Pylyshyn, Z. W. (1994). Some primitive mechanisms of spatial attention. *Cognition*, 50, 363-384.
- Pylyshyn, Z. W. (1996). The study of cognitive architecture. In D. Steier & T. Mitchell (Eds.), *Mind Matters: Contributions to Cognitive Science in honor of Allen Newell*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3), 341-423.
- Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1/2), 127-158.
- Pylyshyn, Z. W. (2002a). Mental Imagery: In search of a theory. *Behavioral and Brain Sciences*, 25(2), 157-237.
- Pylyshyn, Z. W. (2002b). Seeing, acting and knowing: commentary on O'Regan & Noë. *Behavioral and Brain Sciences*, 24(5), 999.
- Pylyshyn, Z. W. (2003). *Seeing and visualizing: It's not what you think*. Cambridge, MA: MIT Press/Bradford Books.
- Pylyshyn, Z. W. (2004). Some puzzling findings in multiple object tracking (MOT): I. Tracking without keeping track of object identities. *Visual Cognition*, 11(7), 801-822.
- Pylyshyn, Z. W. (2006). Some puzzling findings in multiple object tracking (MOT): II. Inhibition of moving nontargets. *Visual Cognition*, 14(2), 175-198.
- Pylyshyn, Z. W. (Ed.). (1987). *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Ablex Publishing.
- Pylyshyn, Z. W., & Annan, V. J. (2006). Dynamics of target selection in multiple object tracking (MOT). *Spatial Vision*, 19(6), 485-504.
- Pylyshyn, Z. W., & Annan, V. J. (2006). Dynamics of target selection in multiple object tracking (MOT). *Spatial Vision*.

- Pylyshyn, Z. W., Burkell, J., Fisher, B., Sears, C., Schmidt, W., & Trick, L. (1994). Multiple parallel access in visual attention. *Canadian Journal of Experimental Psychology*, 48(2), 260-283.
- Pylyshyn, Z. W., & Cohen, J. (1999). Imagined extrapolation of uniform motion is not continuous. *Investigative Ophthalmology and Visual Science*, 40(4), S808.
- Pylyshyn, Z. W., & Dennis, J. L. M. (forthcoming). Can Multiple Object Tracking make use of individual object properties?
- Pylyshyn, Z. W., Elcock, E. W., Marmor, M., & Sander, P. (1978). Explorations in visual-motor spaces. Paper presented at the Proceedings of the Second International Conference of the Canadian Society for Computational Studies of Intelligence, University of Toronto.
- Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, 3(3), 1-19.
- Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA: MIT Press.
- Quine, W. V. O. (1992). *Pursuit of Truth (Revised ed.)*. Cambridge, MA: Harvard University Press.
- Rauschenberger, R. (2004). Attentional capture by auto- and allo-cues. *Psychonomic Bulletin & Review*, 10(4), 814-842.
- Rensink, R. A., & Enns, J. T. (1995). Preemption effects in visual search: Evidence for low-level grouping. *Psychological Review*, 102(1), 101-130.
- Reynolds, R. (1981). Perception of an illusory contour as a function of processing time. *Perception*, 10, 107-115.
- Reynolds, R. I. (1978). The microgenetic development of the Ponzo and Zollner illusions. *Perception and Psychophysics*, 23, 231-236.
- Robertson, I. H., Nico, D., & Hood, B. M. (1995). The intention to act improves unilateral left neglect: two demonstrations. *Neuroreport*, 7(1), 246-248.
- Rock, I. (Ed.). (1997). *Indirect Perception*. Cambridge, MA: MIT Press.
- Rolls, E. T. (1999). Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, 9, 467-480.
- Rosenfeld, A., Hummel, R. A., & Zucker, S. W. (1976). Scene labeling by relaxation operators. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6, 420-433.
- Rosenthal, D. M. (2005). *Consciousness and Mind*. New York, NY: Oxford University Press.
- Rossotti, H. (1983). *Colour: Why the world isn't grey*. Princeton, NJ: Princeton University Press.
- Saarinen, J. (1996). Localization and discrimination of "pop-out" targets. *Vision Research*, 36(2), 313-316.
- Saiki, J. (2003). Feature binding in object-file representations of multiple moving items. *Journal of Vision*, 3(1), 6-21.
- Salinas, E., & Thier, P. (2000). Gain modulation: A major computational principle of the central nervous system. *Neuron*, 27, 15-21.
- Schindler, I., Kerkhoff, G., Karnath, H. O., Keller, I., & Goldenberg, G. (2002). Neck muscle vibration induces lasting recovery in spatial neglect. *Journal of Neurology, Neurosurgery & Psychiatry*, 73(4), 412-419.
- Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology*, 38(2), 259-290.
- Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object: Evidence from target-merging in multiple-object tracking. *Cognition*, 80, 159-177.
- Scholl, B. J., Pylyshyn, Z. W., & Franconeri, S. L. (1999). When are featural and spatiotemporal properties encoded as a result of attentional allocation? *Investigative Ophthalmology & Visual Science*, 40(4), 4195.
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8), 299-308.
- Schulz, T. (1991). A microgenetic study of the Mueller-Lyer illusion. *Perception*, 20(4), 501-512.
- Sears, C. R., & Pylyshyn, Z. W. (2000). Multiple object tracking and attentional processes. *Canadian Journal of Experimental Psychology*, 54(1), 1-14.
- Sekuler, A. B., & Palmer, S. E. (1992). Visual completion of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General*, 121, 95-111.
- Selfridge, O. (1959). *Pandemonium: A paradigm for learning*, Symposium on Mechanization of Thought Processes: National Physical Laboratory Symposium. London: HM Stationery Office.

- Sellars, W. (1956). Empiricism and the Philosophy of Mind. In M. Scriven (Ed.), *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*. (pp. 253-329). Minneapolis, MN: University of Minnesota Press [Reprinted in Sellars, Wilfrid. (1963). *Science, Perception and Reality*, London: Routledge & Kegan Paul].
- Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proc. Natl. Acad. Sci. USA*, 93, 628-633.
- Shapiro, L. (1995). *Affine analysis of image sequences*. New York, NY: Cambridge University Press.
- Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behavioral & Brain Sciences*, 24, 581-601.
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three dimensional objects. *Science*, 171, 701-703.
- Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattention blindness for dynamic events. *Perception*, 28(9), 1059-1074.
- Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, 1, 261-267.
- Simons, D. J., & Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences*, 9(1), 17-20.
- Slemmer, J. A., & Johson, S. P. (2002). Object tracking in ecologically valid occlusion events. Paper presented at the Vision Sciences 2002, Sarasota, FL.
- Slezak, P. (1992). When can images be reinterpreted: Non-chronometric tests of pictorialism, *Proceedings of the Fourteenth Conference of the Cognitive Science Society* (pp. 124-129). Mahwah, NJ, USA: Lawrence Erlbaum Associates.
- Slezak, P. (1995). The 'philosophical' case against visual imagery. In P. Slezak & T. Caelli & R. Clark (Eds.), *Perspective on Cognitive Science: Theories, Experiments and Foundations* (pp. 237-271). Stamford, CT: Ablex.
- Smith, A. D. (2003). *The Problem of Perception*. Cambridge, MA: Harvard University Press.
- Smith, B. C. (1996). *On the Origin of Objects*. Cambridge, MA: MIT Press.
- Snodgrass, M. (2002). Disambiguating conscious and unconscious influences: Do exclusion paradigms demonstrate unconscious perception? *American Journal of Psychology*, 115, 545-580.
- Snyder, L. H. (2000). Coordinate transformations for eye and arm movements in the brain. *Current Opinion in Neurobiology*, 10(6), 747-754.
- Snyder, L. H., Batista, A. P., & Andersen, R. A. (2000). Intention-related activity in the posterior parietal cortex: A review. *Vision Research*, 40(10-12), 1433-1441.
- Snyder, L. H., Grieve, K. L., Brotchie, P., & Andersen, R. A. (1998). Separate body- and world-referenced representations of visual space in parietal cortex. *Nature*, 394(6696), 887-891.
- Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14, 29-56.
- Spence, C., & Driver, J. (Eds.). (2004). *Crossmodal space and crossmodal attention*. New York, NY: Oxford University Press.
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, 74(11, whole No. 498), 1-29.
- Sperling, G. (1967). Successive Approximations to a Model for Short Term Memory. *Acta Psychologica*, 27, 285-292.
- Sperling, G., & Weichselgarter, E. (1995). Episodic theory of the dynamics of spatial attention. *Psychological Review*, 102(3), 503-532.
- Stich, S. (1978). Beliefs and subdoxastic states. *Philosophy of Science*, 45, 499-518.
- Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.
- Stricanne, B., Andersen, R. A., & Mazzoni, P. (1996). Eye-centered, head-centered, and intermediate coding of remembered sound locations in area LIP. *Journal of Neurophysiology*, 76(3), 2071-2076.
- Suganuma, M., & Yokosawa, K. (2002). Is multiple object tracking affected by three-dimensional rigidity? Paper presented at the Vision Sciences Society, Sarasota, FL.
- Swets, J. A. (Ed.). (1964). *Signal Detection and Recognition by Human Observers*. New York: Wiley.
- Tipper, S., Driver, J., & Weaver, B. (1991). Object-centered inhibition of return of visual attention. *Quarterly Journal of Experimental Psychology A*, 43A, 289-298.
- Tipper, S. P., Weaver, B., Jerreat, L. M., & Burak, A. L. (1994). Object-based and environment-based inhibition of return of selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 478-499.



- Titchener, E. B. (1912). The Schema of Introspection. *American Journal of Psychology*, 23, 485-508.
- Tlauka, M., & McKenna, F. P. (1998). Mental imagery yields stimulus-response compatibility. *Acta Psychologica*, 67-79.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189-208.
- Tootell, R. B., Silverman, M. S., Switkes, E., & de Valois, R. L. (1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science*, 218(4575), 902-904.
- Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology*, 40A(2), 201-237.
- Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.
- Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, 29(8), 943-951.
- Tremoulet, P. D., Leslie, A. M., & Hall, D. G. (2000). Infant individuation and identification of objects. *Cognitive Development*, 15(4), 499-522.
- Tresillian, J. R. (1995). Perceptual and cognitive processes in time-to-contact judgements: analysis of prediction-motion and relative judgment tasks. *Perception and Psychophysics*, 57(2), 231-245.
- Trick, L., & Pylyshyn, Z. W. (1993). What enumeration studies tell us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2), 331-351.
- Trick, L. M., Audet, D., & Dales, L. (2003). Age differences in enumerating things that move: Implications for the development of multiple-object tracking. *Memory & Cognition*, 31(8), 1229-1237.
- Trick, L. M., Perl, T., & Sethi, N. (2005). Age-Related Differences in Multiple-Object Tracking. *Journals of Gerontology: Series B: Psychological Sciences & Social Sciences*, 2, 102.
- Trick, L. M., & Pylyshyn, Z. W. (1994a). Cuing and counting: Does the position of the attentional focus affect enumeration? *Visual Cognition*, 1(1), 67-100.
- Trick, L. M., & Pylyshyn, Z. W. (1994b). Why are small and large numbers enumerated differently? A limited capacity preattentive stage in vision. *Psychological Review*, 101(1), 80-102.
- Tsang, E. (1993). *Foundations of constraint satisfaction*. London, UK: Academic Press.
- Tye, M. (1991). *The Imagery Debate*. Cambridge, MA: MIT Press.
- Tyrell, R. A., Rudolph, K. K., Eggers, B. G., & Leibowitz, H. W. (1993). Evidence for the persistence of visual guidance information. *Perception & Psychophysics*, 54(4), 431-438.
- Ullman, S. (1976). Relaxation and constrained optimization by local processes. *Computer Graphics and Image Processing*, 10, 115-125.
- Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.
- Ullman, S. (1984). Visual routines. *Cognition*, 18, 97-159.
- Van de Walle, G., Carey, S., & Prevor, M. (2000). Bases for object individuation in infancy: Evidence from manual search. *Journal of Cognition and Development*, 1, 249-280.
- vanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science*, 14(4), 498-504.
- Viswanathan, L., & Mingolla, E. (1998). Attention in depth: disparity and occlusion cues facilitate multi-element visual tracking (Abstract). *Investigative Ophthalmology and Visual Science*, 39(4), 634.
- Viswanathan, L., & Mingolla, E. (2002). Dynamics of attention in depth: Evidence from multi-element tracking. *Perception*, 31(12), 1415-1437.
- von Holst, E., & Mittelstaedt, H. (1971/1950). The principle of reafference: Interactions between the central nervous system and the peripheral organs. In P. C. Dodwell (Ed.), *Perceptual processing: Stimulus equivalence and pattern recognition* (pp. 41-71). New York, NY: Appleton.
- Voss, P., Lassonde, M., Gougoux, F., Fortin, M., Guillemot, J. P., & Lepore, F. (2004). Early- and late-onset blind individuals show supra-normal auditory abilities in far-space. *Current Biology*, 14(9), 1734-1738.
- Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology*, 45, 205-217.
- Waltz, D. (1975). Understanding Line Drawings of Scenes with Shadows. In P. H. Winston (Ed.), *The Psychology of Computer Vision* (pp. 19-91). New York: McGraw-Hill.
- Warren, D. H. (1970). Intermodality interactions in spatial localization. *Cognitive Psychology*, 1(2), 114-133.
- Washburn, M. F. (1922). Introspection as an objective method. *Psychological Review*, 29, 89-112.

- Watson, D. G., & Humphreys, G. W. (1997). Visual marking: prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological Review*, 104(1), 90-122.
- Watson, D. G., & Humphreys, G. W. (2000). Visual marking: Evidence for inhibition using a probe-dot detection paradigm. *Perception & Psychophysics*, 62(3), 471-481.
- Watt, D. G. D. (1997). Pointing at memorized targets during prolonged microgravity. *Aviation, Space, and Environmental Medicine*, 68(2), 99-103.
- Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
- Wegner, D. M. (2003). The mind's best trick: how we experience conscious will. *Trends in Cognitive Sciences*, 7(2), 65-69.
- Wehner, R., Boyer, M., Loertscher, F., Sommer, S., & Menzi, U. (2006). Ant navigation: one-way routes rather than maps. *Current Biology*, 16(1), 75-79.
- Weiskrantz, L. (1995). Blindsight: Not an island unto itself. *Current Directions in Psychological Science*, 4(5), 146-151.
- Wexler, K., & Cullicover, P. (1980). *Formal Principles of Language Acquisition*. Cambridge, Mass.: MIT Press.
- Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, 68(1), 77-94.
- Whorf, B. L. (1958). *Language, Thought and Reality*. Cambridge, MA: MIT Press.
- Wiggins, D. (1997). Sortal concepts: A reply to Xu. *Mind & Language*, 12(3-4), 413-421.
- Wilson, H. R., Loffler, G., Wilkinson, F., & Thistlethwaite, W. A. (2001). An inverse oblique effect in human vision. *Vision Research*, 41(14), 1749-1753.
- Wilson, J. A., & Robinson, J. O. (1986). The impossibly-twisted Pulfrich pendulum. *Perception*, 15(4), 503-504.
- Wittrich, W. J. (1959). Visual perception and personality. *Scientific American*, 200 (April), 56-75.
- Wong, E., & Mack, A. (1981). Saccadic programming and perceived location. *Acta Psychologica*, 48, 123-131.
- Wright, R. D., Dawson, M. R., & Pylyshyn, Z. W. (1987). Spatio-temporal parameters and the three-dimensionality of apparent motion: Evidence for two types of processing. *Spatial Vision*, 2(4), 263-272.
- Xu, F. (1997). From Lot's wife to a pillar of salt: Evidence that physical object is a sortal concept. *Mind and language*, 12, 365-392.
- Xu, F., Carey, S., & Quint, N. (2004). The emergence of kind-based object individuation in infancy. *Cognitive Psychology*, 49, 155-190.
- Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology*, 24, 295-340.
- Yantis, S., & Johnson, D. N. (1990). Mechanisms of attentional priority. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4), 812-825.
- Yantis, S., & Jones, E. (1991). Mechanisms of attentional selection: temporally modulated priority tags. *Perception and Psychophysics*, 50(2), 166-178.