# ABSTRACT

Title of dissertation: AN ESSAY ON
THE NATURE OF VISUAL PERCEPTION

Ryan Graham Ogilvie, Doctor of Philosophy, 2017

Dissertation directed by: Professor Peter Carruthers
Department of Philosophy

In this dissertation, I address two distinct, but related questions: (1) Is vision encapsulated from higher-level cognitive content? For example, do higher cognitive states like belief and desire alter the contents of vision? (2) What is the scope of visual content? Is the content of vision restricted to "low-level" properties like shape and color or does vision involve a recognitional component? Regarding the first question, I argue that vision is cognitively penetrable, that what we see depends in part on the particularities of our beliefs, expectations, and goals. Regarding the second question, I argue that we visually represent at least some relatively high-level, abstract properties, such as causal interactions, animacy, and facial categories. Both these positions speak to broader issues concerning the epistemic status of our visual capacities. More specifically, we can no longer understand vision as an entirely non-epistemic capacity, one that merely provides us with a structural description of the environment; rather, the visual system carries ontological commitments and by virtue of these commitments it imposes at least a primitive order on what we see.

# An Essay on the Nature of Visual Perception

by

Ryan Graham Ogilvie

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Peter Carruthers, Chair/Advisor
Professor Georges Rey
Professor John Horty
Professor Larry Davis
Dr. Eric Mandelbaum

# Preface

This dissertation is an essay on naturalized epistemology. It attempts to characterize a fragment of our cognitive machinery that makes belief, justification, knowledge, and rationality possible by looking to the empirical sciences of the mind. As such, my aims are *explanatory*. They differ from more traditional epistemological aims of identifying when individuals are justified or rational in holding certain beliefs—what Georges Rey (2014) calls a "working" epistemology. I shall have very little to say about the implications of my view on the more traditional questions that animate "working" epistemologists, though I would be gratified if they found what I have to say useful or illuminating.

The particular fragment that I focus on here concerns the intellectual contributions afforded by the human visual system. Very generally, this project is concerned with the old puzzle of how we come to know to so much about the world given the impoverished and fleeting irritations of our sensory receptors. How does our brain take this input and translate it into stable, coherent, and often veridical representations that guide our actions and inform our deliberations? How are we able to identify food, danger, romantic interest, and the sundry objects of our daily lives? What does perception contribute to these capacities? More specifically my work engages with a view embodied in long-held distinction between sentience and sapience—the idea that vision is a non-epistemic capacity.

When drawing this distinction, contemporary theorists tend to speak of the differences between *perception* and *cognition*. It's unclear why this is the preferred

terminology. (Perhaps because sentience–sapience implies too strict of a division for what is likely a fuzzy border.) At any rate, one problem with using "perception" and "cognition" is that "cognition" is also used as a catch-all term for any kind of brain-related information processing. On this latter usage of "cognition," there is no meaningful distinction between perception and cognition. Thus, I was faced with a number of tradeoffs when choosing my terminology.

In the first chapter, I provide a short history of what I refer to as *sentience–sapience* distinction. For that chapter, I wanted to emphasize the non-epistemic character of historical and contemporary theories of perception. I argue that there's a long history of understanding perception as independent of knowledge forming systems. Moreover, when certain theorist refer to a perception–cognition distinction, this seems to be what they have in mind. But for the remaining chapters, I follow the contemporary use of "perception–cognition" for ease of switching between exposition and argument.

# Foreword

Chapters 2 and 3 of this dissertation are based in part on collaborative work with my advisor, Peter Carruthers. We coauthored "Opening Up Vision: The Case Against Encapsulation" (Ogilvie and Carruthers, 2016b) and a commentary on Firestone and Scholl's (2015) target article in *Behavioral and Brain Sciences* (Ogilvie and Carruthers, 2016a). My committee recognizes that I made substantial contributions to the aspects of the jointly author work as it appears in this dissertation.

# Dedication

*For Dad and Anna.*

# Acknowledgments

This dissertation, quite frankly, would not have been possible without the support and encouragement from my wife, Elizabeth Allan. I am eternally grateful for her love and patience. She is also enterally grateful that I have now completed this dissertation.

I would like to thank my advisor, Professor Peter Carruthers, for many, many helpful comments on previous drafts of this dissertation. There are very few advisors in the field of philosophy that would have tolerated a student delving so deep into the vision science literature. And so I am very grateful and happy that Peter not only tolerated, but encouraged, this way of addressing philosophical questions about the mind.

Georges Rey and I had several lovely meetings in New York over the period in which this dissertation was written. My conversations with Georges were immensely helpful for bringing the broader issues of the empirical literature into focus.

I would also like to thank Eric Mandelbaum for both his insightful comments on early drafts of this dissertation and for graciously accepting to be on my committee. Eric is on faculty at Baruch College in New York City, which means his participation on my committee goes well beyond what professional duty requires. I hope that he at least found the work intellectually stimulating, as a small recompense for his work on this committee.

And finally, I would like to thank my one-year-old daughter, Anna, whose primary responsibility was to provide me with regular breaks from writing.

# Contents

# List of Figures

# Chapter 1: Sentience and Sapience

> *That perceiving and understanding are not identical is therefore obvious; for the former is universal in the animal world, the latter is found in only a small division of it.* (Aristotle, 1984, 427b)

Contemporary theorists have pretty well abandoned Aristotle's theory that humans are unique among the animal kingdom in their capacity to think. But they continue to embrace a related thesis. For Aristotle, the distinction between sentience and sapience is quite literally a distinction between sensing and knowing. Animals can perceive the world, but cannot know things about it. Perception plays a role in explaining the human capacity for knowledge, but it is not constitutive of the ability to cognize.

One can trace a fairly straight line from Aristotle to contemporary thinkers on this issue. Descartes, for example, thought that perceptual processes could be explained in terms of 17$^{\text{th}}$ causal-mechanical interactions. Light enters the eye forming an image on the retina. Post-retinal fibers transmit the image to the central organs of the brain, where upon the image is impressed on the surface of the pineal gland, the seat of "imagination" and the "common sense." (See Figure 1.1.) The scientific challenge of characterizing perception, as Descartes envisioned it, is to

merely describe how the brain transmits the visual image to the intellect. He could not envision an analogous mechanistic story for the capacities required to judge or interpret what the senses depict. Thought is discursive and infinitely expressive, and therefore it could not be a product of a finite physical brain.

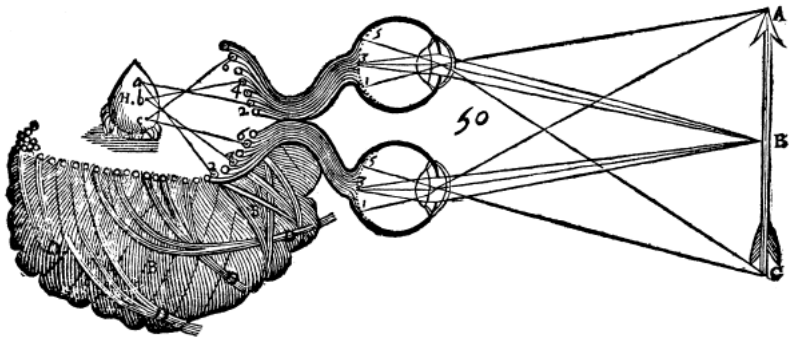


*Figure 1.1: Descartes' model of the human visual system. From* Treatise on Man *(?)*

Contemporary theorists jettison Descartes' dualism, but some continue to embrace the core of the sentience–sapience distinction. Fred Dretske (1969; 2000) distinguishes between what he calls "simple seeing" and "seeing that." The distinction is quite explicitly drawn along the lines of sentience and sapience. Simple seeing is a basic perceptual capacity common to nearly all humans, and shared by a number of other (relatively large and mobile) non-human animals. A corollary of this definition is the fact that simple seeing is independent of our stock of beliefs and dispositions—i.e., simple seeing is a non-epistemic capacity. One can see a pine tree without seeing *that* it is pine tree—i.e., without knowing that it is a pine tree. Indeed, according to Dretske, one can see a pine tree without believing that one is looking at a tree. (At night in a forest, you might mistake a pine tree for shadows cast on a rock face.)

To see *that x* is a pine tree involves forming a post-perceptual judgement about what one sees. This capacity involves the application of a concept—i.e., it involves individuating or categorizing a particular as a pine tree. Those without the concept PINE TREE (e.g., infants) lack the capacity to see *that x* is a pine tree, for any *x*. So for Dretske epistemic seeing goes beyond the perceptual domain, even though it's permissible to speak of *seeing that* as a *perceptual* capacity.

Jerry Fodor (1983) is another contemporary author whose views hew quite closely to the sentience-sapience distinction. His classic formulation of modularity holds that modules are standalone cognitive systems that perform automatic operations on a domain-specific set of inputs. The five sense modalities and the language system are prime candidates for modules.

An important feature of a module is that it is encapsulated from both the outputs of other modules as well as beliefs, expectations, and desires. So, for example, the visual module cannot utilize background knowledge about the environment (that one is in a jungle as opposed to her living room) in the service of constructing a perceptual representation.

Central cognition, by contrast is "anti-modular," and in particular *un*encapsulated. Any belief can, in principle, bear on the confirmation of another. Someone might wonder why the traffic in the New York City is so calm for a Sunday afternoon. Then he might remember that it's the first Sunday of the NFL season and that both the Giants and the Jets are playing: "A-ha! The traffic is calm because nearly everyone is watching the games!" The relationship between beliefs about traffic patterns and beliefs about football games is by no means obvious without quite a

lot of background knowledge. On Fodor's view, this holistic character of central cognition, therefore, plays a central role in explaining our intelligence.

The contrast between modules and central cognition brings the sentience-sapience distinction into stark relief. Modules are quick and automatic (and therefore ecologically beneficial) because they are relatively *unintelligent* systems. Central processes, however, are *intelligent* because they are unencapsulated. It's noteworthy that Fodor (1983; 2008) thinks that genuine intelligence is intractably holistic, and therefore a science of central cognition is unlikely, a claim reminiscent of Descartes' denial that the mind can be explained in terms of physical mechanisms. On the other hand, modules are comparatively unsophisticated, and therefore ripe for scientific investigation.

There are two major themes that come out of this short history of the sentience–sapience distinction that are relevant for understanding the nature of perception. The first is that sentience and sapience are independent capacities. Sentience carries out its operations without any oversight or input from sapience. Knowledge of the world might require sentience, but perception of the world does not constitute a kind of knowledge. The second theme is that sentience is a non-intelligent, non-epistemic capacity. Beings endowed only with sentience lack the ability to know and understand. Psychological systems that underlie sentience, therefore, do not constitute the machinery underlying genuine intelligence.

Regarding the first theme, I argue that vision is cognitively penetrable, that what we see depends in part on the particularities of our beliefs, expectations, and goals. This implies that higher-level systems have some degree of control over what

we see. Regarding the second question, I argue that we visually represent at least some high-level, abstract properties, such as causal interactions, animacy, and facial categories. Both these positions speak to broader issues concerning the epistemic status of our visual capacities. More specifically, we can no longer understand vision as an entirely non-epistemic capacity, one that merely provides us with a structural description of the environment; rather, the visual system carries ontological commitments and by virtue of these commitments it imposes at least a primitive order on what we see.

Chapters 2 and 3 deal with the encapsulation question. In Chapter 2, I address theoretical considerations that motivate the classical modularity theorist to claim that the vision is encapsulated. At best these arguments undermine what I call radical "interactionist" views, views that hold that there no constraints on the information that can be utilized by visual processing. However, the classical modularist's arguments fail to undermine more moderate forms of cognitive penetration. Indeed, in Chapter 3 I present empirical evidence of moderate forms of cognitive penetration. My claims, here, are rather modest primarily because this question has only recently been studied systematically. I suspect that there are other forms of top-down influence on visual processing, but we lack evidence at this point in time.

Chapters 4 through 7 all deal with the issue of perceptual content. Chapter 4 lays the groundwork for the "high–low" debate. All participants in this debate agree that vision at least represents the geometric and chromatic properties of our environments. "Conservatives" hold that vision only represents these properties, while "liberals" hold that vision represents at least some high-level properties, such

as causal interactions or categorical properties (*being a male*). I note that, although conservatism about visual content seems to be the default view among theorists, it's rarely defended explicitly. Clearly, theorists find the view intuitive, but why? I end Chapter 4 by assessing two possible arguments for content conservatism, and argue that they fail to provide a adequate defense of the view.

In Chapter 5, I take up two recent attempts to make progress on the high–low debate. I begin with Susanna Siegel's (2010) phenomenal contrast method. The primary problem with her proposal is that it relies too heavily on introspection and intuition. Because one's intuitions about specific cases seem largely driven by prior theoretical commitments, people with different commitments will come to different conclusions about particular cases. Thus, the phenomenal contrast method fails to provide a method for resolving disputes concerning the scope of visual content.

I then take up a promising, empirically-informed proposal by Ned Block (2014). He argues that facial category (e.g., emotion, race, gender) adaptation provides evidence of high-level visual content. Block argues that since adaptation is generally understood to be a perceptual phenomenon, these high-level attributes must be represented in perception. I argue that while these studies might demonstrate *perceptual* adaptation, Block is wrong to think that these effects are perceptual *because* they are cases of adaptation. I argue that we have good reason to think that at least some cases of adaptation are non-perceptual.

In Chapter 6 and 7, I lay out my positive proposal for why we visually represent high-level properties. Chapter 6 makes the case that "organizing principles" are a fundamental posit of the vision sciences, and that they play an essential role in

explaining phenomena ranging from relatively low-level visual processing (such as Gestalt grouping principles) to object representation. I take these phenomena to be uncontroversial cases of perceptual processing. Thus, they provide a set of cases over which we can make some useful generalizations.

In Chapter 7, I apply these generalizations to purported cases of causal, animacy, and facial category perception. The argument, here, is that since causality, animacy, and facial "perception" all require positing high-level organizing principles (or schemata), and these principles exhibit the hallmarks of visual processing, we ought to conclude that we perceptually represent these high-level properties.

My project in the following chapters is not to advocate abolishing the sentience–sapience. Nor do I deny that there's distinction between perceptual and non-perceptual processes. What I'm calling into question is the idea that the perception–non-perception distinction lines up with the distinction between sentience and *sapience*. The picture of visual perception that emerges from this dissertation suggests that we can no longer understand vision as an entirely non-epistemic capacity, one that merely provides us with a structural description of the environment. Rather, vision is part of human intellectual capacities.

# Chapter 2: The Case for Encapsulated Vision[1]

Classical or "Fodorian" modularity holds that vision is encapsulated from cognition. Over the course of the following two chapters I argue that this is claim is false. The primary aim of this chapter, however, is to give shape to the encapsulation thesis, and to argue that the largely theoretical considerations offered in support of encapsulated visual processing fail to give us good reason to cleave to the view. In Section 2.1, I situate classical modularity within the broader theoretical landscape. In Section 2.2, I discuss the three main theoretical arguments offered in support of encapsulated visual processing and argue that they fail to undermine unencapsulated models. Of course, all participants in this debate agree that it's an empirical question whether vision is encapsulated, not an issue to be addressed entirely by armchair considerations. However, some of the armchair considerations concern the sort of evidence that bears on this issue—whether, for instance, there are plausible confounds for which one needs to control. Thus, it will be useful to describe at the outset some of empirical challenges that one needs to meet in order to provide evidence of genuine unencapsulated visual processes. Section 2.3 walks the reader through the challenges posed by Fodor and Pylyshyn.

[1]This chapter and the following one draw from a piece I co-authored with Peter Carruthers (Ogilvie and Carruthers, 2016b).

## 2.1 Some Geography

When thinking about the issue of cognitive penetration, it's helpful to conceptualize the space of possible views as consisting of a spectrum, from maximal penetration of perception on the left to minimal penetration (or no penetration) of perception on the right. Views occupying the left-most locations on this spectrum hold that perception and cognition are entirely interwoven (see Clark, 2013; Hohwy, 2014; Shea, 2015). These views hold that there is an information processing hierarchy that begins somewhere near the sensory organs (a near universal position in vision science), and proceeds in a highly interactive fashion to higher-level cognitive centers. While the hierarchy is directional, in the sense that higher levels tend to processes more abstract content than lower levels, processing is bi-directional—i.e., processing occurs in a bottom-up and top-down fashion. Maximal interactionists hold that there are few, if any, constraints on the kind of information that can influence lower-level processing.

Fodor (1983) and Pylyshyn (1999; 2003) articulate and defend "classical modularity," a view occupying the right-most location on the spectrum. According to these authors, there's an information processing hierarchy, but perceptual processing, in particular visual processing, takes place within encapsulated modules and operates independently of higher-level cognitive states. That is, higher-level cognitive processes do not interact with lower-level perceptual processes.

The characterizations of these views are intended to be rough glosses of the respective positions, not canonical formulations of them. However, a bird's eye view

of the theoretical geography helps bring a few issues into focus. First, one can see that both views are logically quite strong. The left side of the spectrum says that there are *no* constraints on top-down influence, while the right side says that there are *no* cognitive effects on perception. So both extremes of the spectrum make (strong) negative existential claims.

Second, a vast theoretical landscape lies between the two ends of the spectrum. The view that I defend allows for substantive constraints on when and how higher-level cognitive processes can interact with lower-level perceptual processes. Not all visual processes will necessarily exhibit sensitivity to cognitive processes. And I strongly doubt that any visual process has unrestricted access to all cognitive systems. My objective in the following chapter is to provide evidence of theoretically interesting cases of top-down influences. However, I remain neutral on how just how widespread these sorts of effects are.

Finally, note that the claim that vision is unencapsulated is consistent with weaker forms of modularity (e.g., Carruthers, 2006; Barrett and Kurzban, 2006). The visual system can be functionally distinguished from other perceptual modalities and higher-level cognitive systems without appealing to information encapsulation. Indeed, I claim that the visual system can be understood as a distinct subsystem of the mind/brain. I address this issue in more detail in Chapter 3.5.

My position, then, is rather weak. I merely claim that we find *some* cases of cognition affecting visual processing *some* of the time. Nevertheless, this position is quite controversial. A number of authors argue that the best available evidence points to vision being encapsulated. So let us turn now to the various arguments in

support of the encapsulation thesis.

## 2.2 Positive Evidence

Proponents of classical modularity take a two-pronged approach: (1) they provide positive support, arguing that encapsulated visual processing best fits the data and our theoretical understanding of visual processing; and (2) negative support, arguing that purported cases of cognitive penetration can be accommodated within the classical modularity framework. This section discusses and evaluates the three main positive arguments for classical modularity.

### 2.2.1 Computational Benefits of Encapsulation

The first argument that I consider concerns the computational benefits afforded by informational encapsulation. Encapsulated systems are thought to offer significant speed advantages over unencapsulated systems. And this is a good thing from an evolutionary perspective, as mobile organisms need to swiftly perceive the behaviorally-relevant aspects of its environment if they are to capture prey and evade predation. Thus we should expect intense selection pressure for swift perceptual processing, and this suggests encapsulation. We know, of course, that perception *is* remarkably fast. Using electroencephalography (a measure of electrical brain activity across the surface of the cortex), Thorpe et al. (1996) estimate that it takes humans roughly 150ms to identify whether or not an image contains an animal.

According to Fodor (1983), encapsulated systems are fast because they restrict,

by fiat, which information can be used to arrive at a perceptual interpretation of the incoming sensory signal: "This is to say that speed is purchased for input systems by permitting them to ignore lots of the facts" (p. 70). Note that all (or nearly all) participants in the encapsulation debate agree that the visual system requires some stored information in order to "infer" the distal causes of the impoverished proximal stimulus.[2] Since, Fodor thinks that processing speed is proportional to the size of the search space, the less information to which a system has access, the quicker it will be.

But why think that having access to *less* information will make the problem any more tractable? Given the rather meager sensory input, one might think that having access to more information would be better. Fodor acknowledges that this issue is an empirical one for which he lacks good data. But his case for why we ought to expect modules to be encapsulated doesn't solely depend upon a claim about the relationship between memory and processing speed. Even if search times didn't vary as a function of memory size, systems with access to central cognition would face what he calls the "relevance" problem.

The relevance problem originates, in part, from Fodor's commitment to Quine's (1951) confirmation holism. Fodor thinks that cognition, or "central systems," are Quinean in that any belief can bear on the confirmation or disconfirmation of any

[2]One possible exception to this claim are Gibsonian theorists, who follow Gibson (1984) in rejecting psychological entities as explanatory posits for visual processes. On this sort of view, the visual system does not store information. But as we shall see below, the debate about cognitive penetration boils down to whether there's a *semantic* connection between higher-level cognitive systems and visual processes. Since Gibsonians reject a semantics of vision, there can be no semantic relationship. For all I have said thus far, the Gibsonians might be right, and the debate is a non-starter. But that issue is beyond the scope of this dissertation.

other belief, at least in principle. However, this means that for any piece of information to which the system has access, the system must determine "*how much inductive confirmation it bestows upon the hypothesis*" (Fodor, 1983, p. 71). But determining the extent to which a particular data point bears on the current visual hypothesis is a computation-hungry process, especially for systems with a large number of potentially relevant data points; and while some of these data points could, in principle, bear on the current hypothesis, when facing serious time constraints, for all intents and purposes many are irrelevant: "The point is that in the rush and scramble of panther identification, there are many things I know about panthers whose bearing on the likely pantherhood of the present stimulus *I do not wish to have to consider*... for example, that my grandmother abhors panthers; that every panther bears some distant relation to my Siamese cat Jerrold J..." (p.71). The problem isn't just that large memory stores translates to long search times; it's that for each piece of prior information, one needs to determine its relevance for the current computational task. For memory stores that include a lot of patently non-useful information, the relevance problem is pronounced. Hence, one would expect an architecture that restricts the pool of information to that which is useful for perceptual tasks (whatever, in the end, that might be).

Note that Fodor (1983) also thinks that automaticity explains why classical modules are fast. Automatic (or mandatory) processes are ones that perform their computations independently of other systems. For example, when one looks at a full moon, one can't help but see it as circular. When a native speaker of English (clearly) hears a sentence of English, she cannot help but comprehend it. The mech-

anisms underlying these phenomena are automatic in the sense under discussion.

Automatic processes achieve speed not so much by limiting the amount of information a system receives (although automaticity may entail some *degree* of encapsulation), but by activating a rigid or stereotyped set of rules in response to a specific set of inputs. Here is how Fodor (1983) puts it:

> it may well be that processes of input analysis are fast because they are mandatory. Because these processes are automatic, you save computation (hence time) that would otherwise have to be devoted to deciding whether, and how, they ought to be performed. (p. 64)

It seems then that automaticity can overcome the relevance problem, not by imposing a information firewall, but by imposing stereotyped procedures. That is, a mandatory system doesn't have to "decide" which pieces of information are relevant because there is a well-defined mapping between the inputs and the system's responses.

I raise this alternative explanation for the swiftness of classical modules because it's plausible that a system could be automatic without being encapsulated. Systems can share information without, for example, executive processes mediating how the information is shared. This shows that even on Fodor's view, properties other than encapsulation can plausibly explain fast perceptual processing.

This is an important point to keep in mind, for I shall now argue that neither Fodor nor Pylyshyn can reasonably hold that encapsulation is the only way to achieve swift processing or overcome the computational challenges raised by the

seeming globality of human cognitive abilities. Another way to put this point is that even if we assume that vision (or early vision) is encapsulated, there are at lease some *unencapsulated* processes that are as fast, or nearly as fast, as encapsulated processes; hence, something else must explain the swiftness of these particular unencapsulated processes. But if this right—if there are fast, unencapsulated systems—then one cannot argue that vision must be encapsulated on the grounds that it would otherwise be too slow.

Formulating an argument for this claim is complicated by the lack of consensus about where to draw the line between encapsulated (if there are any) and unencapsulated processes. On the one hand, some think that only early visual areas are encapsulated and that only low-level visual representations are insulated from cognition. On the other hand, some think that encapsulated visual processing produces basic-level recognitional content, such *that* is *dog* or *that* is a *tree*. And the sort of content one thinks marks the border between encapsulated and unencapsulated processing will effect what one thinks is evidence of top-down cognitive effects on perception.

That said, we can avoid this difficulty by noting that the classical modularity theorist faces a dilemma. If the outputs of encapsulated vision are too low level, then they must acknowledge that there are unencapsulated processes that are as fast, or nearly as fast, as encapsulated ones. But if the outputs of vision produce relatively higher-level content, then we find clear evidence of top-down effects. In either case, we must posit fast, *unencapsulated* processes, a result that doesn't bode well for the classical modularity theorist. Let us begin with the first horn of the

dilemma.

Pylyshyn (2003) suggests that the outputs of early, encapsulated visual processing provide a structural description of the immediate environment. A structural description reflects the appearance properties of the scene, and is given primarily in geometric terms (although color and texture would also be specified). Recognition, identification, and belief fixation are all post-early visual processes. As far as Pylyshyn (2007) is concerned, the encapsulation thesis only applies to systems that are responsible for generating a structural description of the environment:

> To determine whether early vision is cognitive penetrable, one needs to factor out functions such as categorization and identification, which require accessing general memory, from functions of early vision, such as individuating objects and computing spatial relations among them, which, by hypothesis, do not. That is why we find, not surprisingly, that some apparently visual tasks are sensitive to what the observer knows, since the identification of a stimulus requires both inferences and access to memory and knowledge (p. 75)

Thus, according to Pylyshyn, evidence of top-down effects on object *recognition* does not count as evidence of cognitive penetration of early vision because these processes utilize inference and memory—processes beyond the scope of the impenetrability thesis.

If this is where the encapsulated–unencapsulated line is drawn, then there must be at least some swift but unencapsulated processes. From a evolutionary

perspective, it doesn't matter whether one is able to quickly process the shape and color of an object if one cannot also rapidly apprehend its behavioral significance. Merely representing the low-level features of a snake, for example, doesn't tell us whether it is dangerous. Of course, we may be "hard wired" to have certain affective responses to snake-shaped objects—a kind of "snake-shape-detector" module that triggers increased levels of vigilance and the formation of motor plans. But positing a relationship between snake-shaped objects and behavioral responses affirms my general point: merely representing the geometric properties of the environment is of no use to the organism unless the geometric properties are *also* appropriately connected to behavior guiding systems. Hence, insofar as one thinks that there is an evolutionary imperative for swift processing of early vision, one should equally think that there is an evolutionary imperative for swift recognition and identification processes.

Indeed, this is what one finds when the issue is studied empirically. A study by Grill-Spector and Kanwisher (2005) shows that categorizing objects at a "basic level" yields the same level of error and occurs just as quickly as detecting the presence of an object. Detection, here, is understood as seeing (though not necessarily recognizing) an object. Thus, we can identify detection with a structural representation of the object/image.[3] In the main experiment of this study, the researchers ran three different types of trials: (1) "detection" trials, where participants are asked

[3]Taxonomies are typically hierarchically organized, from the very specific (e.g., *Drosophila melanogaster*) at the lowest level to the very general (e.g., *thing*) at the highest and most abstract. A "psychologically basic" level of categorization is one that is maximally informative, in the sense that it maximizes both generality and differences across category boundaries. Standard examples of basic level categories are *fly*, *dog*, *cat*, *house*, and *car* (Rosch et al., 1976).

to identify whether they see an object or visual texture; (2) "categorization" trials, where participants are asked to name the "basic-level" category (e.g., bird, dog, fish, flower, house); and (3) "identification" trials, where participants are asked to name the objects at a "subordinate level" (e.g., Harrision Ford, pigeon, German shepherd, shark). All trials have the same basic structure. Participants are first presented with a scrambled (unintelligible) image, and then briefly presented with the "target image" of a building, face, musical instrument, etc. for varying durations (17, 33, 50, 68, or 167ms). (The duration of the scrambled image also varies, so that one cannot predict the onset of the target image.) Grill-Spector and Kanwisher (2005) find that the error rate for detection and basic level categorization is identical across the different exposure durations, while error rate for the identification trials is substantially greater for all exposure durations (except the 167ms exposure duration, where detection, categorization, identification were perfect).

In a similarly designed experiment, they measured response times and performance, and found that performance and accuracy for detection and were nearly identical. This experiment uses the same basic trial structure as the previous experiment, though this time participants are forced to choose between two options. Detection involves choosing between an "object" and a "texture," categorization involves choosing between, for example, "car" and "not car," and identification involves identifying whether the image is an instance of a specified category (e.g., jeep). In line with the first experiment, the "identifying trials" generated greater error and slower response times than the other two trial types.

Grill-Spector and Kanwhisher (2005) argue that the parallel performance for

detection and categorization undermines the widely held view that seeing an object occurs prior to object recognition, suggesting that detection and categorization of basic level categories are the same process.[4] Whether or not one accepts that inference, these results pose a serious problem for Pylyshyn's view. If it's true that only early vision is encapsulated, and that the outputs of this system provide a structural description of the environment, then something else besides encapsulation must explain the speed with which object recognition occurs.

Of course, one might deny that encapsulated processes are restricted to providing a low-level structural representation. Indeed, Pylyshyn is admittedly quite speculative in his description of the outputs of early vision. Fodor (1983), for example, suggests that encapsulated vision outputs representations that correspond to basic level categories. Indeed, Mandelbaum (2016) argues that Grill-Spector and Kanwhisher's (2005) findings provide positive support for the idea that an encapsulated visual module outputs basic-level content. This position would also avoid the behavioral relevance problem facing Pylyshyn's proposal. If one is looking at panther, the visual system might deliver something like [BIG CAT, THERE]. Here the behaviorally relevant content is processed quickly by the encapsulated system.

This response effectively involves expanding the representational capacities of the visual system. Attempting to characterize the outputs of vision is notoriously difficult, in large part because there are very few uncontested facts to constrain one's theory. So expanding the representational capacity of the visual system is a prima

[4]Further support for their conclusion comes from the finding that when the shorter exposure durations generated detection errors for a particular image, they also generated categorization errors, and vice versa. That is, detection and recognition error rates were correlated for particular image-duration pairings.

facie legitimate move to make when confronted with Grill-Spector and Kanwhisher's (2005) results. This move, however, leads to the second horn of the dilemma.

It's well known that stimuli are more readily identified if they are situated within a "meaningful" or congruent context. For example, if the context is a kitchen, people can recognize a loaf of bread (congruent stimulus) more easily than a drum (incongruent stimulus) (Palmer, 1975). This suggests that expectations and beliefs about one's current environment influence the recognition of these objects. Note, however, that BREAD and DRUM are likely cases of basic level categories. Thus, it seems that the representation of contextual cues, what is considered high-level semantic information, can modulate basic level categorization processes. That is, if perception outputs basic level representations, then it seems that vision is unencapsulation.

My concern isn't just that there's a single study that seems to undermine the idea that basic level categories are processed in encapsulated systems. There is a wealth of evidence showing that semantic priming and context can facilitate object recognition (Bar, 2004; Biederman, 1972; Davenport and Potter, 2004; Potter, 1975; Rémy et al., 2013). And it's precisely this sort of evidence that Pylyshyn (2003) in the above passage dismisses as *apparent* cases of cognitive penetration. Thus, the modularity theorist can't consistently maintain the claim that basic level content is the product of encapsulated processes and allow (as the data shows) that recognition is influenced by top-down factors.

One might wonder if the recognition processes in the Kanwisher and Grill-Spector study are the same recognition processes at work in the priming studies.

Perhaps semantic priming only affects unencapsulated, post-perceptual processes, leaving (slow) encapsulated recognition untouched by top-down effects. The problem with this objection is that context and priming sensitive recognition occurs at similar time scales to those found by Kanwisher and Grill-Spector. (See Rémy et al., 2013.)

But even if I'm wrong that basic-level categorization is sensitive to cognitive primes, the classical modularist faces a different problem that arises from the Kanwisher and Grill-Spector study. It turns out that the response times for basic-level categories (e.g., *dog*) and sub-ordinate categories (e.g., *poodle*) are *both remarkably fast*, differing by less than 100ms. So even if we lacked evidence of top-down effects for basic-level categorization, other levels of category recognition exhibit remarkable speed. If the claim is that only basic-level categorization is encapsulated, then we have evidence of a speed advantage, but only a small one. But now it's unclear whether such a small speed advantage is consistent with the original rationale for positing encapsulated visual processing. Is the claim now: vision must be encapsulated because otherwise visual processing would be 100ms slower? This position seems implausible in large part because it's now unclear whether the speed difference is actually due to encapsulation and not some other factor, such as category type.

Let us take stock of the argument so far. If the output of encapsulated systems is a structural description of the visual scene (and post-early visual systems are unencapsulated), then it looks like at least some unencapsulated processes must be fast. This is because (a) a structural description lacks the relevant ecological information necessary for survival, and (b) basic-category recognition is fast. This is

the first horn of the dilemma. Now, if the modularity theorist claims that encapsulated vision delivers behaviorally relevant, basic level representations, then his view is inconsistent with a large body of data that shows that this kind of information is unencapsulated. This is the second horn of the dilemma. An easy way out of the dilemma is to simply give up on the cognitive impenetrability thesis.

I'd like to draw a more general point from this discussion of the supposed computational benefits of encapsulation. If the problem with unencapsulated perceptual processing is that it would be prohibitively slow (because it had to determine the relevance of every panther-related belief that one might hold), it seems that the problem is just as acute for belief fixation. Suppose I'm about to leave my house to catch a bus. But I want to know if there's a chance the bus will arrive earlier than is posted in the schedule. If I think there's a good chance it'll be early, I'll wear a shirt that doesn't need ironing. But there's no way I'd ever catch the bus—indeed, no way I'd do much of anything—if I had to consult every bus-related belief in order to form a belief about its arrival time. This is no way for any cognitive system to do business.

While I think there's something correct about the Quinean intuition that, in principle, any belief can bear on the confirmation of any other, I think it is pretty clearly not how the brain works in real time. The computational intractability of the relevance problem shows that something must constrain which information is available to a process at a given time; however, it doesn't entail that the constraint must consist of an impenetrable firewall.

Carruthers (2006) makes a similar point: "The Fodorian argument from com-

putational tractability, then, does warrant the claim that the mind should be constructed entirely out of systems that are frugal; but it doesn't warrant a claim of encapsulation, as traditionally understood. . . " (p. 59). Systems can be frugal (i.e., meet the holistic computational challenge), he argues, by employing search and decision heuristics which function to restrict the amount of information available to a system at various stages of computation. Although much more could be said here about how the implementation of heuristics "solves" the relevance and frame problems, this is not my burden. To the extent that the relevance problem is a genuine theoretical problem, the human mind serves as an existence proof that it is solvable. Indeed, it solves it admirably, for much of ordinary, humdrum belief fixation occurs remarkably fast.

Fodor might respond to this argument as follows. Yes, it's true that unencapsulated processes can be fast, and there has to be something other than encapsulation which explains this fact. Nevertheless, what explains, at least in part, why perceptual processes are fast is the fact that they are encapsulated. That is, it remains quite plausible that vision is encapsulated because encapsulation greatly reduces the complexity of the computational problem, and therefore produces a speed advantage. This response suggests that encapsulation can nonetheless play an important role in explaining the speed of visual processing. I now want to offer a tentative argument that undermines this claim.

Suppose that the visual system is entirely encapsulated, and is faced with the particularly thorny problem of deciding whether some luminance contrast is a change in surface reflectance properties or a shadow. As it turns out, almost

any paradigmatic visual cue could help resolve this problem. When one looks at standard results in visual psychology, one finds that all sorts of cues figure into the computation of surface properties. For example: depth information can be used (does the entire visual area reside on the same plane of depth?); color information can be used (is there a difference in color between the two surfaces?); global features of the scene can be used (are there any global contours that would suggest that we are dealing with a change of surface as opposed to a shadow?); texture information can be used (is there a texture difference between the two areas?).

Note that each of the visual properties is determined by a collection of individual cues. Consider depth cues: motion parallax, kinetic depth cues, perspective, relative size, aerial perspective, accommodation, occlusion, curvilinear perspective, texture gradients, lighting/shading, stereopsis, and convergence. The received view amongst psychologists is that the visual system (somehow) integrates multiple depth cues to form an overall determination of relative depth. Each of these separate cues can be informative, but aren't always. Furthermore, it's well known that auditory information is integrated into visual processing. The auditory case is nice for the point that I'm making, precisely because it doesn't seem particularly relevant for disambiguating surface properties. Yet, it would not be shocking if psychologists found that, under certain circumstances, auditory information is used to disambiguate shadow from surface. When one looks at the literature on visual processing, one finds a staggering number of variables that can and often do bear on whether a particular contrast change is perceived as a change in surface properties or change in illumination. What I'm claiming—i.e., what I think the vision science shows—

is that visual processing is, to an interesting degree, holistic. This claim might strike some as entirely implausible, so I should probably explain what I mean by "interesting degree."

The kind of holism that Fodor has in mind is confirmation holism, in particular Quinian confirmation holism. Quinean holism holds that any belief about the world can, in principle, bear on the confirmation of any other belief. Global holism quantifies over an individual's entire belief set. And, indeed, when philosophers think about confirmation holism they typically have global holism in mind. But we can think of holism occurring locally, where the set of data points that bear on a hypothesis are a subset of all possible data points. Suppose, contra Quine, that physics and psychology are completely independent; facts about subatomic particles have, in principle, no bearing on facts about mental states. (E.g., perhaps dualism is true.) It could still be true that the body of knowledge concerning the physical is holistic in the sense that no critical experiment can confirm or disconfirm a particular theory or hypothesis *within physics*.

So if I'm correct that visual processing is non-globally holistic in character, then the relevance problem applies. To the extent that it is a problem, it's a problem for any system of sufficient complexity. Another way of putting this point is as follows: the relevance problem is about holism, not about the extent to which a system is unencapsulated. For even on Fodor's view central systems are a degenerate example of encapsulated systems; my central system is encapsulated from your central system.

This is a very general point about the relationship between encapsulation and

holism. But, again, nothing about this last argument shows that the human vision system is unencapsulated. This is an empirical claim that I will make by appealing to work in cognitive neuroscience in Chapter 3. What I have been trying to show in this section, however, is that the theoretical considerations about the computational benefits of encapsulated processing are not particularly compelling.

### 2.2.2 The Epistemic Benefits of Encapsulation

I now want to address the argument for the epistemic benefits of encapsulation. Like the argument for the computational benefits of encapsulation, the worries are largely based on how perceptual systems would operate if they weren't encapsulated. Here is how Fodor (1983) describes the worry:

> Pylyshyn's point is that a condition for the reliability of perception, at least for a fallible organism, is that it generally sees what is there, not what it wants or expects to be there. Organisms that don't do so become deceased. (p. 68)

It seems that Fodor and Pylyshyn think that *if* perceptual processes were sensitive to background beliefs and conative attitudes, organisms would be prone see what they expect to see (even when those expectations are false) and see what they want to see (even if those desires are unfulfilled). The worry, then, is that because unencapsulated perceptual systems are unreliable, they would not have been apt for survival, and therefore would not have been selected.

Given the way Fodor and Pylyshyn characterize the dire consequences of unen-

capsulated perceptual processing, they seem to assume that expectations and desire would *determine* the content of perceptual states. They seem to assume, for example, that if I desire a fully stocked refrigerator, then that's what I'll see when I peer inside—even if all it contains is condiments and a few half empty takeout containers. If cognitive penetration entailed this sort of top-down dominance of visual processing, then I agree that cognitive penetration would be epistemically pernicious.

However, I see no reason for why a defender of cognitive penetration would be committed to such a strong view. Cognition can modulate visual processing without dominating those processes. While some theorists have made strong claims about the theory ladenness of visual perception (e.g., Hanson, 1965; Churchland, 1979), I argue that the effects of cognition on perception are far more subtle. I shall say more about the ways in which cognition affects visual processing in the next chapter. But on the present issue, note that we can turn Fodor and Pylyshyn's argument on its head, as there are good prima facie reasons for thinking that the effects of higher-level cognitive states on visual processing can be epistemically beneficial.

One area of research suggests that an organism's knowledge about it's current environment can constrain possible interpretations of the scene when the sensory signal is especially noisy or ambiguous (Yuille and Kersten, 2006; Summerfield and Egner, 2009; Alink et al., 2010; Hegdé and Kersten, 2010; Pinto et al., 2015). When the sensory signal contains above average noise or is especially ambiguous, relying solely on information or mechanisms contained within the visual system would both increase the likelihood of error and slow down the time it takes to generate an interpre-

tation. But if cognitive systems can supply information to the visual system—either as beliefs about the particular environment context in which the organism is located, or as goals the organism is pursuing—this information can be used to home in on a likely interpretation of the visual scene. Indeed, recent experimental evidence suggests a positive correlation between the amount of noise or uncertainty in the sensory input and the amount of top-down involvement in perceptual processing (Hsieh et al., 2010). From a Bayesian perspective this makes sense. The greater the uncertainty assigned to the incoming sensory signal, the greater weight background information is given when forming a perceptual hypothesis.

Of course, having access to faulty background information or applying it in inappropriate situations can lead an organism astray. A study by Payne (2001) shows that racial stereotypes can increase misidentification of objects. White participants were primed with an image of either a Black or White face and then presented with a degraded image of either a gun or a "crime-neutral" tool (e.g., pliers, socket wrench, drill). Payne found that, under time pressure, priming participants with Black faces increased error rates for the crime-neutral tool—i.e., White participants mistook crime-neutral tools for guns. The prevailing theory that explains these results is that White participants are particularly susceptible to stereotypes of Black individuals being associated with gun-related crime. Under time pressure, this implicit bias primes White participants to expect a gun in the degraded image.

Do the participants literally see a gun when in fact they are presented with a pair of pliers? Do implicit biases affect genuine perceptual processing or is this just a post-perceptual effect? This is a difficult question to answer. But it doesn't

much matter for present purposes. For whether or not Payne is measuring a genuine perceptual effect or a post-perceptual bias, the epistemic upshot is the same: one's background theory or implicit biases can lead one astray.

But notice that this observation undermines Fodor and Pylyshyn's original argument that access to background beliefs and desires would be epistemically pernicious. If the epistemic consequences of biases are the same, irrespective of the locus of the effect, then there's no special problem about the prospect of beliefs "infecting" visual processing. Or, at any rate, Fodor and Pylyshyn need to explain why the problem isn't just as bad in the cognitive case. Without such an additional argument, I don't see any reason why epistemic concerns should count against unencapsulated models of visual processing.

### 2.2.3 The Persistence of Illusions

The final positive argument for encapsulation that I want to discuss concerns the appeal to the persistence of visual illusions. For the last century or so, vision scientists have discovered a number of displays that generate inaccurate (or illusory) percepts that seem insensitive to background beliefs. The Müller–Lyer illusion in Figure 2.1 is a standard example. The top line appears longer than the bottom line, even though the lines are of the same length. Those familiar with the illusion are quite aware that the lines are of the same length, yet the illusion persists. A number of authors claim that this is a problem for unecapsulated theories of visual processing (Fodor, 1983; Firestone and Scholl, 2014; Pylyshyn, 1999). They argue

that since we have high confidence in the belief that the lines are the same length, we should expect that this belief to extinguish the illusion. That is, if beliefs and expectations inform visual processing, then we ought to expect cognition to correct the visual error.

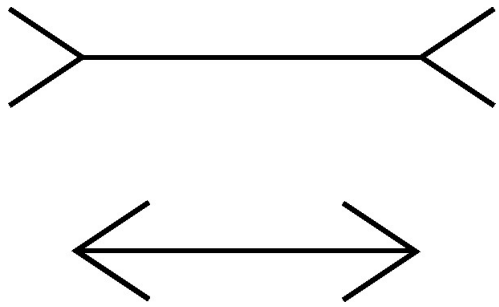

*Figure 2.1: The Müller–Lyer Illusion. The top line appears longer than the bottom even though the two lines are of identical length. (Original drawing.)*

My response to this argument is two-fold. First, even if illusions are entirely resistant to background beliefs and expectations, this doesn't show that vision must be encapsulated. A moderate view may hold that cognitive penetration occurs, yet claim one or another of the following:

(i) Some perceptual processes are encapsulated from the influences of certain beliefs or expectations, while other aren't.

(ii) Certain kinds of background knowledge (in principle) cannot alter certain perceptual processes (but other kinds can).

(iii) There exists a regulative mechanism that determines when and how background knowledge can affect visual processing.

(iv) All visual processing can (in principle) be biased by cognitive factors, but the

extent to which beliefs affect visual processing depends on the amount of, and strength, of the top-down evidence.

Thus, the mere fact that visual illusions persist despite conflicting background beliefs doesn't show that visual processing must be encapsulated. So in order for the persistence of illusions to count as evidence for classical modularity, it must be the case that encapsulated visual processing must provide the best explanation of the phenomenon. In the following chapter, I provide empirical evidence of cognitive penetration. If I'm correct that cognitive penetration exists, then complete encapsulation is false.

Nevertheless, the persistence of illusions calls out for an explanation, and so what, besides encapsulation, might explain this phenomenon? Item (iii) offers a possible explanation. As I note in the previous section, recent Bayesian-inspired theories of visual processing posit that top-down signals are thought to play an important role in constraining possible interpretations of the scene when the bottom-up signal is particularly noisy or ambiguous. We might expect, then, that top-down contributions will be minimized when the bottom-up visual processing is relatively unambiguous and noise-free. Indeed, when we turn to Bayesian explanations of illusions, we find reason to think that the visual processing that gives rise to the inaccurate percept is neither especially noisy nor ambiguous.

Howe and Purves (2005) argue that the Müller–Lyer illusion is explained in terms of prior assumptions about environmental regularities, also known as "natural scene statistics." For contours that subtend the same angle on the retina, the visual

system "assumes" that those with adornments closer to the center of the line tend to be closer (and therefore shorter) than those with adornments further towards the ends of the lines. Howe and Purves (2005) note prior work that shows that "arrow fins" are not necessary in order to generate the illusion. Square- or circle-shaped adornments will produce the same illusory percept. They argue that it's the relative locations of the adornments on the line that matter. By analyzing a large set of natural scene images, these authors were able to identify Müller–Lyer-like contour configurations within the images, and found that centrally-adorned contours correlated with objectively shorter environmental contours than did end-adorned contours.

Thus, in most naturalistic circumstances, this assumption yields accurate size estimates. It's only when this assumption is applied to a contrived stimulus that the assumption leads to an inaccurate percept (see Howe and Purves, 2005; Teufel et al., 2013; Weiss et al., 2002, for the same point). But from the perspective of the visual system, there is nothing particularly noisy or ambiguous about the sensory signal. The visual system utilizes relative depth and size cues and settles upon an interpretation of the visual scene. Higher-level systems monitoring noise and error levels are being told that everything is in order: there is no need for further processing.[5] So one possible explanation for the persistence of illusions is that the signals produced when viewing an "illusory display" are sufficiently free of noise and ambiguity, and therefore don't call upon higher-level beliefs or expectations to constrain the

[5]Note that the claim that illusions are the result of normal visual processing in the presence of a contrived display is not unique to Bayesian accounts of visual processing. Antony (2011), for example, makes the same point from a classical computational standpoint.

interpretation of the sensory signal.[6]

Much more needs to be said about how all this works; but recent work in computational neuroscience offers at least some reason to think that we don't need to posit encapsulated visual processes in order to account for the persistence of illusion. The main point that I have been making in this section is that the persistence of illusions is one data point among many and that any successful theory will have to account for this phenomenon, but it's no means obvious that unencapsulated models cannot account for this data.

## 2.3 Negative Evidence

It should be noted that neither Fodor nor Pylyshyn think that the previously discussed arguments conclusively establish that vision is encapsulated. They are well aware they are working in the field of psychology, where the data is noisy and the theories are somewhat sketchy. Thus it's important that they undermine the alleged evidence in support of unencapsulated or interactionist models of visual processing.

There are two main ways that modularists attempt to undermine unencapsulated models of visual processing. The first strategy involves undermining purported

[6]Notice that this explanation sketch presupposes a distinction between the way "visual" assumptions affect visual processing and the way high-level assumptions affect visual processing. The visual assumptions underlying the Müller–Lyer illusion are applied in an automatic or mandatory fashion—i.e., the visual system *always* (or almost always) treats Müller–Lyer-like configurations as depth and size cues, irrespective of judgments in other systems. But higher-level assumptions are applied in a context sensitive manner and require that a mechanism monitor and control when they should be applied. One might see this as evidence that the visual system is distinct from higher-level cognitive systems. Indeed, I would agree with this assessment. But we shouldn't confuse classical modularity (which entails encapsulation) with weaker conceptions of modularity. I think we can distinguish the visual system from other systems; I just deny the visual system is entirely encapsulated.

evidence of cognitive penetration. This is an important contribution to the debate, for as Fodor and Pylyshyn point out, much of the evidence for cognitive penetration is ambiguous between the two hypotheses. I refer to the second strategy as "Pylyshyn's how-possibly challenge": if cognition trades in concepts and perception doesn't, it's unclear how abstract, amodal content could possibly interact with low-level stimulus parameters in a semantically sensitive manner.

### 2.3.1 Accommodating "Top-Down" Effects in Classical Modularity

Both Fodor and Pylyshyn bring a number of subtle issues to bear on the putative evidence for cognitive penetration. Indeed, I take this to be one of their most valuable contributions to the debate, as many of the experimental confounds they discuss continue to be relevant in contemporary debates. For example, in a review of the literature purporting to demonstrate top-down effects on vision, Firestone and Scholl (2015) argue that many experimenters fail to control for *intra-modular effects* (top-down effects that occur within the visual system), *post-perceptual effects* (effects at the cognitive level), or *attentional selection effects* (effects due to attentional mechanisms selecting the inputs to visual processing, but leaving visual processes untouched).

Before discussing these confounds in greater detail, let us look at Pylyshyn's (1999) original characterization of cognitive penetration to see why they are genuine confounds:

> For present purposes it is enough to say that if a system is cognitively

> penetrable then the function it computes is sensitive, in a semantically coherent way, to the organism's goals and beliefs, that is, it can be altered in a way that bears some logical relation to what the person knows. (p. 343)

There are two broad features of this characterization that deserve emphasis. The first is that Pylyshyn requires that there be a semantic connection between the high-level and lower-level states. In other words, not just any old top-down causal factor can count as cognitive penetration. And this makes sense, given that causal relations are relatively cheap to come by. For example, Macpherson (2012) discusses an imaginary (but plausible) case where a student believes that the next day's exam will be difficult. The stress triggered by this belief causes a migraine, which in turn produces "scintillating scotomata"—the illusory perception of flashing lights. In this case, it seems, the student's belief *caused* a change in perceptual experience (or visual processing), and so one might be tempted to characterize this as a case of cognitive penetration. However, because any number of different belief contents could cause similar levels of stress (e.g., the belief that tomorrow's big race will be tough, etc.) which would cause a migraine, this case fails to satisfy Pylyshyn's semantic constraint.

The second feature of Pylyshyn's description of cognitive penetration that deserves emphasis is that the cause of the top-down effects need to be genuinely cognitive. Fodor and Pylyshyn both endorse top-down processing within modular systems. But the "high-level" processes in these cases are not instances of genuine

belief or desire. Since this point sets up some of the architectural distinctions that figure into the the different potential confounds, let us turn to those now.

*Intra-Modular Effects*: Both Fodor and Pylyshyn explicitly allow for top-down effects to occur *within* a system. Visual processing occurs in stages, and later stages of processing can feedback to earlier stages. But because top-down effects of this sort originate from within the visual system, they are not genuinely cognitive top-down effects.

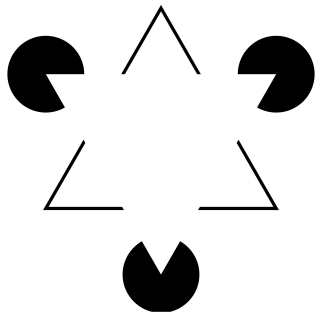

*Figure 2.2: The Kanizsa Triangle, an example of illusory contour completion. (Permission granted under the Creative Commons Attribution-Share Alike 3.0 Unported licence.)*

There are a variety of effects that can be explained in this way, but one plausible case is illusory contour completion. When one looks at the Kanizsa Triangle (see Figure 2.2), one typically sees illusory contours linking the three "packmen." What looks to be going on in this case is that global contour information (the edges of the packmen mouths) influences local contour processing, effectively "filling in" a contour where there is no luminance contrast in the distal or proximal stimulus. One explanation of this phenomenon is that higher-level visual processing areas, sensitive to the global co-linearity of the "packmen mouths," cause neurons in lower

areas to fire in the absence of a contour. This would be a case of a top-down effect, but one that occurs within the visual system. Thus, it is not a case of cognitive penetration.

*Inter-Modular Effects*: Cross-modal interaction occurs when one sensory system modulates activity in another, and is often cited as evidence against classical modularity. The McGurk Effect is a standard case of cross-modal interaction. For example, when the phoneme /ba/ is overdubbed onto a video recording of someone saying /ga/ people reliably hear the intermediate phoneme /da/. It seems, then, that visual information can alter auditory information. While this sort of evidence might undermine other aspects of classical modularity (e.g., see Prinz, 2006), it doesn't provide evidence of cognitive penetration. Nothing about cross-modal effects implicates higher-level cognitive systems. However, if "cross-modal" interactions are mediated by higher-level cognitive systems—i.e., an auditory signal triggers a memory, which modulates visual processing—then this would count as cognitive penetration.

*Post-"Early Vision" Effects*: As note above, Pylyshyn conceptually distinguishes between processes that analyze the incoming sensory signal and processes that identify or classify the stimulus. It's only the former processes that are encapsulated; the latter processes can be and often are influenced by beliefs, expectations, and desires. The fact that recognition occurs so swiftly and is susceptible to cognitive priming puts pressure on the idea that there is a clean distinction between so-called sensory analysis and recognition. Nevertheless, if one wants to make the case that vision is penetrated by cognition, one needs to control for post-perceptual biases,

and independently establish that the locus of the effect is perceptual in nature.

*Attentional Effects*: Pylyshyn (1999; 2003) argues that, although cognitively guided attention is a top-down effect on perception, it violates the semantic constraint. His argument for why attention should not be considered a form of cognitive penetration is driven largely by looking at cases of spatial attention. According to Pylyshyn, spatial attention works by pre-perceptually selecting inputs for preferential processing. While I agree that spatial attention is not a case of cognitive penetration, the issue gets more complicated when we look at other forms of attention, such as feature based attention. I shall address these issues in more detail in the following chapter. Nevertheless, Pylyshyn raises some legitimate concerns that the effects of spatial attention can be confused for cognitive penetration, and it's instructive to see why this form of attention fails to satisfy Pylyshyn's semantic constraint.

One way attention can select visual inputs is by guiding eye movements. If I overtly shift my gaze (i.e., by moving my eyes) to the top left-hand corner of the screen, I have effectively altered which inputs my visual system receives, therefore altering visual processing. While direction of gaze is often guided by explicit goals (I want to look at *this* side of the drawer for my favorite socks), this case clearly fails to exhibit a semantic relationship between the high-level cognitive states and the effect on visual processing, as the only role that cognition plays is to inform motor systems of the location of my socks. It's only in virtue of a difference in the environment that I see something different. If I were looking at a entirely homogenous scene, looking here or there wouldn't alter visual processing. Moreover, the same goal (to

look towards *that* side of my visual field) could produce any number of different effects on visual processing depending on the environment in which I am located.[7]

According to Pylyshyn, something analogous happens when we covertly shift attention (i.e., when one shifts attention without moving one's eyes or head). On his model of covert focal attention, cognitive systems send signals carrying spatial content (e.g., coordinates) to the visual system, which then gets used to single out regions of space for preferential processing. Again, these sorts of cases seem to violate the semantic constraint on cognitive penetration. I have a belief about the location of my socks; but the belief merely informs the visual system about which region of space to devote resources, as opposed to altering the content of the perceptual state.

Pylyshyn (2003) further argues that a number of purported cases of cognitive penetration can be accomodated within the classical modularity framework when one is sensitive to the effects of attention. The two main cases of purported cognitive penetration he discusses concern images containing ambiguous contours and "expert perception."

Two-tone images (also known as "Mooney" images) are high contrast, black and white images that initially appear as random ink splotches until the viewer identifies the overall form embedded in the image. The famous Dalmatian dog image (by R.C. James) is a prime example. Theorists typically note two features about this kind of image. First, once one sees the dog, it seems impossible to "unsee" it. Second, telling a naive viewer to look for a dog (and if that doesn't

[7]Note, also, that the causal relationship in this case, as in Macpherson's migraine example, is too indirect. The belief about the location of the socks, in conjunction with the goal to find the socks, informs the visuo-motor systems where to look. But the causal relationship is mediated by non-mental, external factors—viz., the features in the environment.

work, to look for a Dalmatian!) seems to speed up recognition of the embedded shape. Proponents of cognitive penetration often cite these observations as evidence for their view (e.g., Churchland, 1998), arguing that one's background knowledge primes the visual system to complete the shape. In cases where one has already identified the Dalmatian, or where an individual is told to look for a dog, the viewer uses conceptual knowledge to extract the implicit form embedded in the image.

Pylyshyn offers an alternative theory. Because attention can affect where we spatially attend, knowledge that there's a Dalmatian in the image (as well as beliefs about normal image composition) can be used to guide one's attention to the relevant locations in the image. For example, attention might be directed towards locations where a head might be, facilitating contour completion in that area. Once the initial contours are complete for the head, the rest of the contours can be readily identified. As Pylyshyn notes, there is often a significant delay between giving someone a hint about the object embedded in the image and the point at which the object perceptually "coheres." It's during this time that Pylyshyn thinks that participants are performing a search:

> What may be going on in the time it takes to reach perceptual closure on these figures may be simply the search for a locus at which to apply the independent visual process. This search, rather than the perceptual process itself, may thus be the process that is sensitive to collateral information. (p. 80)

So while knowledge about the image has an effect on what one perceives, on Pylyshyn's

model, that knowledge merely plays an instrumental role of guiding one's attention to the relevant areas of the image.

Pylyshyn appeals to similar considerations in explaining the perceptual expertise exhibited by radiologists, chicken "sexers," atheletes, artists, and so on. What explains why a radiologist can see a spiral fracture in a noisy x-ray image, and lay observers cannot? Clearly she has learned skeletal anatomy and what particular landmarks look like in an x-ray, which on his view is entirely post-perceptual. The radiologist can use this sort of information to to efficiently and effectively deploy spatial attention towards the relevant areas of the image. But as far as the radiologist's visual system is concerned, it operates in more of less the same manner as any one else's.

Whether or not Pylyshyn's explanations are entirely correct, they are plausible enough to seriously question whether these kinds of cases require positing unencapsulated visual processing. Thus, one needs to exercise caution when interpreting the empirical literature. In order to conclude that a particular top-down effect is a case of cognitive penetration, one needs to rule out spatial attention as a competing hypothesis.

### 2.3.2 The How-Possibly Challenge

Pylyshyn asks if cognitive penetration were to exist, what sort of mechanism would we need to posit? One possible mechanism could involve "proto-hypotheses" operating as a kind of filter, which increases the system's sensitivity to a particular

feature of the visual field. For example, if I am on a beach, I might expect to see sailboats on the water, and this expectation might increase my sensitivity to sail-shaped objects on the horizon. If something like this occurs, argues Pylyshyn (1999), "we need a mechanism that can be tuned to, or which can somehow be made to select a certain sub-region of the parameter space" (p. 353). But Pylyshyn doesn't think this sort of mechanism is possible:

> Unfortunately, regions in some parameter space do not in general specify the type of categories we are interested in—that is, categories to which the visual system is supposed to be sensitized, according to the cognitive penetrability view of vision. (p. 353)

What Pylyshyn is saying is that in order for any type of filtering to operate on low-level stimulus properties, such as light intensities, contour detection, motion, color (i.e., "regions of a parameter space"), some collection of parameter values would need to encode information about the typical shape of sailboats. Presumably Pylyshyn thinks that this is necessary in order for my concept of SAILBOAT to trigger a proto-hypothesis. But as he rightly notes, a mere collection of sensory representations do not, by themselves, encode this kind of information. So if cognition penetration perception, how is this possible?

It's unclear how much weight we ought to put on this argument. Clearly it's unfair to expect the proponent of cognitive penetration to provide a detailed explanation of how it works, as detailed models in any area of psychology are, at present, nonexistent. Moreover, one of the deepest theoretical challenges facing psychology

concerns the interface between perception and cognition. We lack agreement on where the border is and how it should be drawn. And one of the great mysteries of the mind is how we manage to conceptualize the world around us—i.e., how, from fleeting sensory registrations, we manage to categorize the world into discrete places, properties, and events.

However, this challenge is an important one. One reason is that doing so requires that we build and test models against the empirical data. In the following chapter, I discuss a number of models that begin to elucidate how higher-level processes interact with visual processing. While much more needs to be said about how cognitive states interact with visual processing, they nonetheless offer a glimpse of how higher-level cognitive states modulate visual processing. In Chapters 6 and 7, I argue that the psychological structure underlying visual processing involves much more than simple sensory parameters. The picture that emerges from this dissertation is one where we find increasingly abstract organizing principles within the visual system, principles that impose structure on the incoming visual signal. These organizing principles are primitive forms of categorization, a bridge between the geometric analysis of the light array and conceptualized thought.

# Chapter 3: Vision Unencapsulated

In the previous chapter I argued that the positive and negative considerations in favor of encapsulation are not convincing. In this chapter, I provide empirical evidence of cognitive penetration. But before I do this, I want to address Pylyshyn's idiosyncratic view of attention, as one's views on attention influences how one evaluates evidence for cognitive penetration.

## 3.1 Pylyshyn on Attention

Pylyshyn (2007) claims that attention operates at two "loci" in the information processing hierarchy:

> First, [attention] might operate on the input to the visual system, by enhancing and/or attenuating certain properties of the input... Second, it might operate by enhancing and/or attenuating the availability of certain perceptual categories—what Bruner meant by "perceptual readiness"—the ready availability of categories of perception. (p. 89)

Although attention can enhance or attenuate visual inputs, Pylyshyn is quite adamant that attention never affects processes within early vision: "My position is that attention operates at both loci but not in between, i.e., not within the early-vision

system itself" (p 90). Recall that, for Pylyshyn, early vision outputs a structural description of the environment, so visual processing within early vision involves processing information about contours, surface layout, figure-ground segmentation, depth, motion, etc.

Pylyshyn's discussion presupposes a tripartite architecture: a pre-early vision input selection stage, early vision, and post-early visual processing. Indeed, there is good reason to think selective attention operates prior to the early visual processes. For example, the thalamus—a subcortical structure along the visual pathway—is a good candidate for an attentional selection stage. Saalmann et al. (2012) argue that cortical projections to the thalamus filter or gate signals originating from the eyes. However, it's far from obvious that the only way for attention to modulate early visual processes is by altering the pre-perceptual input selection stage.

Consider a study by Carrasco et al. (2004) that found that covertly attending (versus not attending) to a contrast grating increases its apparent contrast. When participants' attention was directed (via a symbolic visual cue) to one of the contrast patches on the display, the contrast of the attended patch required less contrast to appear the same as the unattended patch. So it appears that attention boosts the gain of contrast processing for feature within the focus of attention, thus suggesting that attention *directly* alters early visual processing. Indeed, this is the main conclusion that Carrasco and colleagues draw from their work, noting that it appears "as if attention boosted the actual stimulus contrast" (Carrasco et al., 2004, p. 312). And in a review of the last 25 years of attention research Carrasco (2011) writes that "the confluence of psychophysical, single-unit recording, neuroimaging studies, and

computational models" indicate "that attention modulates early vision" (p. 1486).

If this is correct—and we have good reason to believe it is—we have cognitive systems directly affecting visual processing, thus undermining Pylyshyn's claim that attention never directly alters early visual processing. However, the mere fact that attention directly modulates visual processing doesn't entail that we have a case of cognitive penetration—as least not if we hew close to Pylyshyn's discussion of the concept. This is because we need not posit a *semantic relationship* in all cases where attention directly modulates visual processing.

Recall the study by Carrasco and colleagues. In that experiment, we don't need to posit a semantic relationship between higher-level goal state ("look towards *that* spot") and the changes in visual processing, viz., the boost in contrast sensitivity. While the goal state must communicate to visual processing centers which region of visual space should be singled out for preferential processing, once the spatial content has been communicated, the visual system can carry out its operations in a generic fashion—i.e., in the same way that it would if attention was directed by something other than the particular goal (e.g., by an automatic, bottom-up visual cue). In other words, the high-level state tells the visual system which regions of space to preferentially process, but it doesn't tell the visual system how to perform its computations. This explanation is consistent with encapsulated visual processing. Of course, we have just looked at a single case of attention, and so by no means have we demonstrated that other forms of attention don't count as cognitive penetration.

It's now well established that attention is not only spatially based, but also object- and feature-based (see Scholl, 2001, for a review). For example, attention

can spread across rather large areas of the visual field (much larger than standard instances of spatial attention) so long as one is attending to a single object or feature. Explaining how particular objects or features are singled out by higher-level states for preferential processing may require positing a semantic relation between cognition and perception. How our visual systems parse a scene may depend on how we attend; and this may, in turn, depend on our current goals and beliefs. Note that Pylyshyn has made important contributions to establishing that attention is directed towards objects. However, because he denies that attention ever affects visual processing, he never considers the possibility that feature-based or object-based attention might meet his definition of cognitive penetration.

Of course, the mere fact that attention *could* modulate vision in a way that exploits a semantic relation between cognition and perception shouldn't convince anyone that, in fact, it does. I shall argue below that some cases of cognitive penetration are plausibly interpreted as cases of attention. It may well be true that some of the top-down effects I discuss are attentional in nature. But I will also argue that there is a semantic relation between the higher and lower processes. Hence, the fact that a top-down effect is attentional doesn't entail that it isn't genuine cognitive penetration. I raise this point, in part, to pre-empt criticisms that the studies I review can easily be interpreted as demonstrating the effects of attention.

## 3.2 Goal Directed Effects in Visual Processing

One prima facie reason to think that visual processing might be affected by goals and desires is that extracting and making sense of information in the light array is computationally and neurologically expensive. Objects have a multitude of visual properties and even small areas of the visual field can contain an abundance of visual detail. Simultaneously extracting all the worldly detail that *could be extracted* by the visual system is not likely possible. So it makes sense that some visual features receive preferential processing from the visual system, and that which features are selected is likely to be task sensitive. Suppose that you're looking for a golf ball in the tall grass along the edge of the fairway. We might think that this goal could alter visual processing in such a way as to facilitate the search. For example, the goal of looking for a little white ball might boost neural responsiveness for small radius contours (golf-balled sized) and high luminance and spectral contrasts (white against the green), as well as dampen responsiveness for large radius contours, intersecting contours, and low luminance and spectral contrasts (the many blades of grass). Plausibly, this would allow for a more effective and computationally economical search.

So is there any evidence for task- or goal-related effects on visual processing? There's now a growing body of literature in the cognitive neuroscience of vision that looks at these sorts of effects. Some of this research looks at how action planning enhances the processing of action-relevant visual features. For example, Gutteling et al. (2015) find that preparing a grasping action enhances object size discrimination,

when compared against preparing a pointing action. The thought is that preparing a grasping action demands more precise spatial information about the object than is required for pointing, and therefore a non-visual source of information (e.g., motor plans or intentions) modulates how visual processes extract spatial information from the scene.

To test this hypothesis, the authors looked at fMRI activation patterns in early visual areas (V1) involved in encoding low-level spatial properties, and found that they were able to predict with 70 per cent accuracy whether participants were preparing to point to, or grasp, an oblong shape. In other words, even though the bottom-up visual information was the same in both the grasping and pointing conditions, visual processing in in V1 differed between the two conditions in a predictable fashion. This suggests that the intention to grasp (as opposed to point) modulates how early visual processing encodes visual information from the scene. Of course, this evidence is, at best, suggestive that goals can modulate visual processing; it's by no means conclusive.

We find stronger evidence that goals and intentions influence early visual processing in a set of studies led by Charles Gilbert (Li et al., 2004; Ramalingam et al., 2013).[1] Using single cell recordings from early visual areas of macaque monkeys, Li et al. (2004) find that primary visual cortex (V1) encodes different spatial information from identical stimuli when the macaques are performing different visual discrimination tasks. The researchers began by training the macaques to perform

[1]For ease of exposition, I shall just refer to the Li et al. (2004) study, the main findings of which were confirmed by Ramalingam et al. (2013).

two independent tasks on two distinct sets of stimuli.

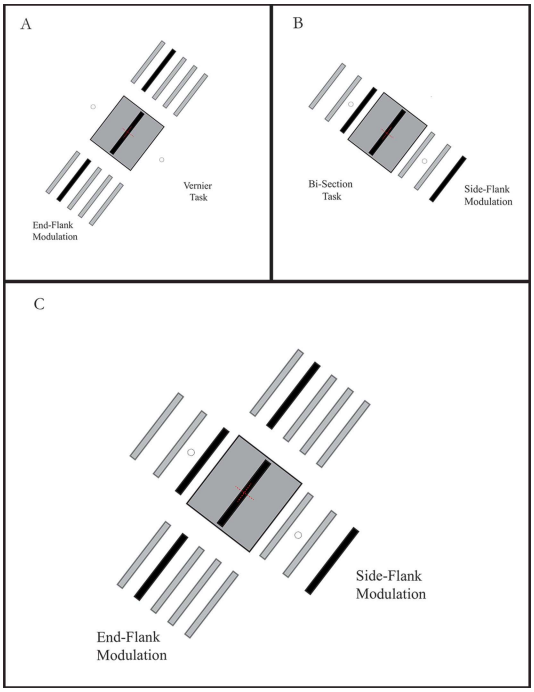


*Figure 3.1: Figures A and B show the individual displays for the vernier and bi-section tasks, respectively. C represents the composite stimuli used in the main experiment. In the actual display, the bars were either white or green against a black background. When, for example the end-flanking bars were green, this indicated that the vernier task was to be performed. Mutatis mutandis, for when the end-flanking bars were green. (Original drawing based on descriptions provided in (2004)*

The stimuli for the *vernier task* consist of three bars placed (roughly) end-to-end. The two end bars (always collinear) are located in five possible positions (two offset right, two offset left, and one inline with the middle bar). (See Figure A in 3.1.) The macaques fixate on a central cross until the display appears, at which time they orient their eyes to a fixation dot located on the side on which the end bars are offset. The stimuli for the *bi-section task* consist of identical bars, this time placed side-by-side, with the two side flanking bars located at varying distances from the center bar, again with five possible configurations. (See Figure B in 3.1.) The

bi-section task is essentially the same as the vernier task, except that the macaques fixate to the side on which the side-flanking bar is closest to the middle bar.

For the main experiment, Li and colleagues use hybrid stimuli, constructed by combining the vernier and bi-section stimuli, for a total of 25 possible stimuli (figure C in 3.1). This allows for two possible tasks to be performed on a single stimulus. The experimenters use colored bars to cue which of the two tasks (the vernier or bi-section) the monkey is to perform. During the main experiments, the researchers recorded activity from individual neurons with receptive fields that correspond to the region of visual space where the stimuli are presented.[2]

The reason Li and colleagues are interested in neural responses from V1 is because this area is specialized for the encoding of local contrast gradients (edges), orientation, spatial frequency, color, and binocular disparity (i.e., standard low-level visual features). It remains an open theoretical question as to just how visual information is processed by this area of the brain, but it is an uncontroversial early stage in cortical visual processing. Thus, if visual processing is encapsulated from high-level goals, then we should expect no difference across the two task conditions. That is, if visual processing is bottom-up, taking only input from the eyes, then we should find identical activation in response to identical stimuli, independent of task.

This, however is not what Li and colleagues found. Instead, they found that

[2]The surface of visual cortex is retinotopically organized. Very roughly, neighboring neurons in the retina project to neighboring neurons along the surface of visual cortical areas, from what is called a feature map. There are several distinct feature maps in the visual parts of the brain. Because of this organization, individual neurons in cortex have receptive fields (small regions of visual space that they can "see"), regions in which the cortical neurons maximally fire in response to their "preferred" stimulus type (lines at particular orientations, motion, visual texture, particular colors, etc.) Li and colleagues selected edge sensitive neurons with receptive fields that corresponded to various regions of the display.

neural responses varied considerably depending on which task the macaque was performing. The standard way of modeling single cell neural data is to construct a tuning curve, a graph that plots firing rates across a stimulus dimension. Li and colleagues constructed two sets of tuning curve, one where end flanks were manipulated and the other when the side flanks were manipulated. In each set, they plotted two tuning curves for each neuron, one for the task-relevant condition and the other task-irrelevant condition. For example, in the tuning curves where the end flanks are manipulated, they constructed curves based on the firing rates when the macaque was performing the vernier task (the task-relevant condition) and they constructed curves when the macaque was performing the bi-section task (the task-irrelevant condition).

The experimenters found that, on the whole, cells were more sensitive (showed greater changes in activation) to changes in the location of the side and end flanks when those changes were relevant for the visual task being performed. For example, when the experimenters manipulated the side flanks while the macaques were performing the vernier task (the task-irrelevant condition), neural response curves tended to remain flat—i.e., neural responses across the different side-flank positions remained (more or less) constant. However, when the experimenters manipulated the end-flanking bars while the macaques were performing the vernier task (the task relevant condition), neural responses varied considerably across the different positions of the end-flanking bars. They found a parallel result when they manipulated the side-flanking bars: manipulating the side-flanking bars produces a differential neural response only when the monkeys were performing the bi-section task.

As a way to quantify these differences between task-relevant and task-irrelevant conditions, Li and colleagues utilized a mutual-information metric, which involves using patterns in the neural recording data to predict likelihood of the stimulus configuration. They found that task-relevant neural responses better predicted the stimulus than did task-irrelevant ones. Thus, in addition to the differential neural responses across task relevant and irrelevant condition, the difference in mutual information suggests that the neural activation patterns are predictive of the task-relevant stimuli. And this suggests that these populations are encoding task-relevant configural information, which would be useful for the monkey when performing the visual task.

A plausible explanation for this data is that the neural population in V1 encodes different spatial relationships amongst the elements depending on the task. When a monkey switches tasks, its visual system selects a different perceptual grouping of the elements, producing a Gestalt "flip," such as when one flips the orientation of the Necker cube. Performing the vernier task requires grouping the collinear elements; thus, collinear spatial relationships are coded more precisely than the relationship amongst the parallel elements. A similar explanation would hold when the bi-section task is performed.

This explanation sketch converges with previous findings that show how scene context influences how local visual properties are processed. For instance, global contour features can induce "illusory" contours, as in the "modal" contour completion of the Kanizsa triangle (Gilbert and Wiesel, 1992). Li et al. (2004) suggest that these types of "contextual" effects can be explained by 'horizontal" connections

within the retinotopically-organized contour maps within the visual system. Very roughly, local brain areas process local features in the visual field. A "horizontal" or lateral network connects these local cortical areas, such that activity in one area can modulate the activity in another. It's thought that global co-linearity modulates local responses to contrast gradients in neighboring receptive fields. The upshot of this kind of cortical architecture is that contextual or global activity systematically alters local computations, which in turn changes how visual information is encoded at a more global level. Gilbert and colleagues hypothesize that top-down connections use this horizontal network to modulate feature processing in a task-dependent manner.

Admittedly, this explanation requires further development and empirical confirmation, particularly the claim that the "horizontal" network is modulated by cognitive factors. One might, therefore, wonder whether the results merely reflect an intra-modular top-down effect, as opposed to a genuine case of cognitive penetration. In the abstract, intra-modular top-down effects occur when information, which is stored or processed later in the visual processing hierarchy, is fed back to earlier areas. My earlier discussion of modal contour completion is a good example of an intra-modular top-down effect. Later stages of visual processing analyze global properties of the scene, which in turn modulate earlier stages of processing.

How might Li et al.'s (2004) findings be reinterpreted as an intra-modular effect? It's not obvious how this might arise, given that the visual context is the same across trials. However, a possible alternative explanation for these results might go as follows. Recall that the macaque's task cue consisted in colored bars. So

if they were to perform the bi-section task, the middle and side-flanking bars would be green. Because the recorded neurons were color insensitive, and the colored bars were isoluminant to the remaining bars, one might think that this wouldn't be a confound. However, other visual areas might use color as a Gestalt grouping cue (grouping green objects as part of a larger whole), and these higher-level color areas could modulate the activity of the recorded neurons (thereby making them more sensitive to the spatial relations amongst the colored bars and not the others). If this explanation is correct, we'd have an intra-modular effect on the processing of spatial relations, as opposed to a top-down effect.

Li and colleagues anticipated this worry, and ran a control experiment, whereby the monkeys performed a simple bi-section task on three side-by-side bars. There were two conditions: one where the bars were all green, and another where the center bar was green and the side-flanking bars were white. If color was driving the effect in the original experiment, then one would expect there to be a difference between the side-flank tuning curves. This is not what they found. Instead, they found identical tuning curves for side-flank manipulation for each stimulus condition. So it appears that the color of the stimulus elements did not drive the modulation of the neural responses.

A further potential objection is that this effect could be explained in terms of spatial attention, and hence is not a genuine top-down effect. Li and colleagues looked at the effects of attention using this paradigm. The monkeys were instructed to perform the tasks on either the original stimulus, located within the recorded receptive field (the attending-to condition), or an alternative stimulus located in

the opposite visual hemisphere (the attending-away condition). They found that when the monkeys were attending away from the array of bars within the receptive field of the recorded neurons, the curves were flat regardless of the task being performed—i.e., the "task" effect was extinguished. This makes sense, as the monkeys were performing the tasks on stimuli located outside of the receptive fields of the recorded neurons. But when the macaques attended to the stimuli within the recorded receptive fields, the same differential response curves from the main experiment reappeared.

Does this mean that the effect is merely the result of attention? Well, *spatial* attention does seem to play a role in increasing sensitivity to the spatial relationships. Indeed, spatial attention appears to be necessary for the effect found in Li et al.'s (2004) study. However, spatial attention alone fails to explain the data. Recall the rationale for why spatial attention shouldn't be considered genuine cognitive penetration. Higher-level cognitive centers direct the focus of attention to a particular location in visual space. But once the location is identified, purely generic attentional mechanisms do whatever it is that they do (e.g., increase the signal to noise ratio for contours). But in the Li et al. study, merely positing a "generic" attentional operation won't do. For if there were a generic attentional operation, we would expect it to produce the same modulation to identical stimuli located in the same region of visual space. But this is not what the main experiment shows. What Li and colleagues found was a differential response to identical stimuli under identical attentional conditions.

So it appears that we have good reason to think that behavioral task does have

some influence on how visual processing operates. It's implausible to think that the effect is merely a cross-modal one, as there are no non-visual sensory cues that correlate with the task. And even if the effect is the result of feature-based attention, Pylyshyn's semantic criterion is satisfied. Thus, this experiment provides good reason to think that goals or desires affect low-level visual processing. However, this is just one study or paradigm. If the aim is to reject encapsulation as a fundamental feature of visual processing, more than a single case is required.

## 3.3 Expectation Driven Effects

There have been a number of studies of late looking at the effects of expectations on visual processing (Jiang et al., 2013; Kok et al., 2012a,b; St. John-Saaltink et al., 2015; Summerfield and de Lange, 2014). Many of these studies rely on fMRI data, which taken alone is not particularly informative. However, when paired with well-established psychophysical paradigms, basic neuroscience, and biologically inspired computational models, fMRI data can place substantive constraints on psychological explanation. Thus, what we are looking for with fMRI data is convergence with existing knowledge of how the brain is organized, how it processes information, and standard behavioral data.

We find this sort of convergence in Kok et al. (2014), who find that, in the absence of visual input, expecting a particular stimulus produces similar feature-specific patterns in V1 to those produced while actually viewing the stimulus. Furthermore, they found that the same expectations facilitated perceptual discrimina-

tion, which strongly suggests that expecting a stimulus increases the visual system's sensitivity in accordance with the expectation. In other words, we have good reason to think that expecting a stimulus of a particular orientation initiates the corresponding orientation-selective processing in early visual areas, and that this priming effect facilitates visual discrimination.

In the behavioral experiment, participants viewed a sequence of two contrast gratings. The first was always orientated at either 45° or 135°, the second grating was tilted a few degrees clockwise or counterclockwise from the first. The task was to identify whether the second grating was rotated clockwise or counterclockwise from the first. Immediately preceding the first grating, a predictive auditory cue would sound. Prior to the trial, participants were *explicitly* told that the cues, consisting of either a high- or low-frequency tone, would predict with 100% validity the orientation of the subsequent contrast grating. In 75% of the trials, the cue was genuinely predictive of the orientation of the subsequent grating. In the remaining trials, the prediction was violated. The question was whether there would be a difference in participants' ability to discern the direction of rotation of the second grating, measured in terms of response times. As predicted, participants' performance was faster when their expectations were met, as opposed to when they were violated.

The behavioral component of the fMRI experiment consisted of the same basic task. In 75% of trials a valid predictive auditory cue was followed by a series of two orientation gratings. The task was the same as in the behavioral experiment, though it was not what they were measuring. In the remaining trials, participants heard one of the two auditory cues, but the orientation gratings were omitted. Thus,

although participants expected to see a grating at a particular orientation, one did not appear.

As expected, Kok and colleagues found that when a grating of a particular orientation was present (e.g., 45°), voxels sensitive to 45° orientations were more active than the voxels sensitive to 135° orientations.[3] This merely confirmed that the fMRI was tracking cells sensitive to the particular grating orientations. The main finding, however, was that even when the gratings were omitted, they continued to find orientation selective activation. In trials with the 45° cue, they found selective activation of 45°-sensitive voxels, and in trials with the 135° cue, they found selective activation of 135°-sensitive voxels.

Taken alone the behavioral evidence is entirely neutral as to the locus of the effect. The speedier response times in the non-violation trials could be due to an increase in post-perceptual "readiness"—i.e., a priming of a post perceptual judgment. However, the fact that early visual areas showed an orientation-sensitive response clearly favors the hypothesis that the differences in response times are due to top-down effects modulating visual processing. That is, when a person hears a low tone (and they are told low tones predict 45° grating), cells that are preferentially sensitive to 45° gratings increase their firing rate.

Note that one could consistently accept Kok at al.'s (2014) fMRI results and yet maintain that the behavioral results are due to a post-perceptual judgment bias.

[3]Voxels are the smallest unit of volume in which fMRI can reliably identify changes in oxygen levels in the blood. The actual size of a voxel can differ for different applications and brain areas. During an independent phase of the imaging session, Kok and colleagues used generic contrast gratings to identify which voxels were preferentially sensitive to 45° and those preferentially sensitive to 135°. This allowed them to independently determine the orientation selectivity of each voxel.

But such a position seems highly unmotivated. Recall that the experimenters found orientation sensitivity when stimulus gratings were present. And this method of using fMRI to identify orientation sensitive voxels has been independently established (Haynes and Rees, 2005; Kamitani and Tong, 2005; Norman et al., 2006). Indeed, it's relatively uncontroversial to find orientation-selective activation in these areas using these methods. The remarkable finding is that the researchers find orientation selective activation when participants merely expect a grating of a particular orientation—i.e., in the *absence* of a stimulus. Since we find orientation-selective priming in early visual areas, we have good reason to think that the improved reaction time in the behavioral component of the experiment is the result of perceptual processing. But could this effect be the result of attention, or even a crossmodal effect? Let's start with the possibility of this being a crossmodal modulation of visual processing.

Recall that an auditory cue indicates whether the participants should expect a 45° or a 135° grating. One might think that the auditory signal directly modulates the visual processing—i.e., without any cognitive mediation. This hypothesis seems doubtful, however. This is because the tone-orientation pairings were entirely arbitrary and varied across trials. Thus, there's nothing about the tone or the way it is processed that could explain the *orientation selective* responses in early visual areas. Moreover, in order for the cues to be predictive, participants need to be explicitly informed of the tone-orientation pairing at the beginning of each trial. That is, participants needed to interpret the linguistic information and put it to use in the behavioral and fMRI trials. So it's implausible that these results could be explained

solely in terms of a crossmodal effect.

The possibility that the improved response times and the underlying changes to visual processing are the result of attention raises a number of subtle issues concerning the nature of attention. One might think that what is going on in this situation is that one's expectation causes one to attend to the expected orientation. Directing one's attention to a particular orientation has the consequence of priming the visual system to process contours at that orientation. Essentially, we are talking about a form of feature-based attention, where the feature is contours of a particular orientation. For example, when looking at a neo-classical façade, I can attend to a particular location on the façade (the middle), or I can attend to pattern running along the frieze. The latter form of attention would be feature-based, because attention is directed towards some feature of the façade (the frieze), as opposed to a particular location.

One question that arises is how this is supposed to work, particularly on Pylyshyn's view. Recall that for Pylyshyn, attention works by "selecting" particular visual features or objects for preferential processing. According to this view, bottom up processing parses the scene into proto-objects and features. Feature-based attentional mechanisms operate by selecting which features should be singled out for further processing. But recall that Kok and colleagues show activation of orientation sensitive processing in the absence of a stimulus. In this case, there are no features to be parsed preattentively, and thus, there is nothing to select.

Pylyshyn's (2007) model of attention is quite speculative, however, and other theorists might reject his model, but nonetheless think that the Kok et al. results

show an attentional effect, as opposed to a genuine case of cognitive penetration. For instance, one might think that attention works by priming or enhancing visual processing for the attended features. So even if no stimulus is presented, one might nonetheless find feature-specific changes to visual processing. Indeed, it's entirely possible that feature-based attention is at work in this case, as the behavioral tasks involved identifying the orientation of stimuli—i.e., the task seems to involve attending to the orientation of the stimuli.

The primary thing to recognize is that, even if one thinks that the priming of orientation sensitive visual areas is the result of attention, one must accept that there is some sort of causal relationship between the expectation and the modulated visual state. (There's got to be some way for the expectation to alter the visual processing, either semantically or merely causally.) The view that I'm targeting, then, holds that the relationship lacks the appropriate semantic connection. But how is this supposed to work? In the case of spatial attention, it's fairly easy to see how a non-semantic causal relationship holds between an expectation and visual processing. Here, one forms an intention to look towards a particular location, and sends a top-down signal carrying information about where to attend. But once this occurs, independent mechanisms carry out operations on the visual input (boost contrast or suppress noise in a particular region of visual space).

However, there seems to be something rather different going on in the Kok et al. (2014) case. The content of the expectation produces an analogous effect in the visual processing stream. Participants who were told that a high-frequency tone predicts a 45° contrast grating exhibited activity in areas selective for 45° orientations.

Furthermore, the cue-orientation pairings are entirely arbitrary and vary across trials, and therefore the causal connection can't be a brute learned association. In this case those defending encapsulated visual processing need to explain the correlation between the expected orientation and the activation of orientation-specific neurons, and it's not obvious how this could be done without positing a semantic relationship.

Interestingly, in the condition when the stimulus was omitted, the visual activation of orientation-sensitive neurons was in same retinotopic location (and no other) as when the stimulus was present. In other words, expecting a stimulus of a particular orientation did not activate orientation-sensitive processing for receptive fields outside of where the stimuli were presented. This suggests that spatial attention plays a role in modulating how high-level expectation influence visual processing. Not only must one expect an orientation in order to produce this sort of priming effect; but one must expect an orientation at a particular location. That is, one must be attending to a particular location in order to generate this effect. Again, this might be a case of feature-based attention. But from the fact that some top-down effect is an attentional effect, we can't infer that it's not a case of genuine cognitive penetration.

Let us pause for a moment and take stock. When presenting alleged evidence of cognitive penetration, my primary burden is to show that the effect is the result of a cognitive processes. In the Kok et al. (2014) study, I have been arguing that full-blown cognitive expectations are driving (a) the increase of behavioral performance and (b) the orientation-selective activation in the absence of stimuli. My first argument gives reason to think these effects are not due neither to direct

auditory-visual connections (without cognitive mediation). My second argument concerns attention: even if this phenomenon is a case of feature-based attention, the evidence points to a semantic connection between the expectation (i.e., expecting a particular orientation) and the behavioral and fMRI data. For if there were no semantic relation, it's hard to see why we would get the orientation-specific effects from merely hearing an auditory tone.

We find further support for the idea that expectations modulate visual processing in Kok et al. (2013). Participants in this experiment were given the task of judging the direction of semi-coherent motion of a "random-dot-motion pattern." (In such stimuli most of the dots move randomly, but a subset move coherently in a single direction.) They were told that the direction of motion could be anywhere within a 90° arc. In fact, however, the directions of motion were restricted to five equally-spaced directions within a 90° arc. During both the experimental and prior learning trials, orientation-predictive auditory cues were presented.[4] Participants were told to ignore the tones, and for the most part learning was implicit.[5] The task involved participants manipulating a comparator line to match the perceived orientation of the motion. This task was was performed in both a conventional behavioral setting and in an fMRI scanner.

[4]The auditory-orientation pairings worked as follows. For each participant, a high and low tone were paired with either motion at 27.5° or motion at 62.5° . Each tone was predictive 60% of the time. So if the pairings were High–27.5° and Low–62.5° , on 60% of the trials when the participant heard a high tone, the screen would display motion at 27.5° 60% of the time. On the remaining 40% of high/low-tone trials, the auditory cues were equally assigned to the remaining four orientations.

[5]Post-experiment interviews revealed that 80% of participants suspected no relationship between the auditory cue and orientation of motion. Of the remaining 20%, one participant was aware of the true significance of the cues, one was aware of a relationship, but had their predictive character reversed; and the remaining three participants suspected a relationship between just one of the auditory cues and presented orientation.

Behaviorally, Kok and colleagues found that auditory cues biased participants' judgments of orientation in the cued direction. For example, when hearing a tone that predicted motion oriented at 27.5°, a participant might perceive motion objectively orientated at 45° as oriented at 35°. This tells us that the participants had learned the statistical associations between the tones and the motion direction, and that their implicit expectations were biasing their judgements. For our purposes, the question is whether this was merely an effect on post-perceptual judgment, or whether these expectations were influencing motion-processing early in the visual system itself.

The fMRI data confirm the latter. The investigators used a "forward-modelling" approach to estimate the perceived direction of motion on each trial. This essentially involved collecting fMRI data from motion-selective voxels in V1, V2, and V3 on each trial, and using this data as input to a direction-encoding artificial neural network. They predicted that if the locus of bias was in visual processing, then this should be revealed in the fMRI data, and the model should produce biased estimates of the perceived orientation. And this is what they found. The forward models better matched the participants' reports of the perceived direction than they did the actual directions of motion. Moreover, when the researchers found a positive correlation between the bias produced by an individual's model and the individual's perceptual bias. For example, if someone showed a stronger bias than others in the behavioral condition, then so did her fMRI forward model. This provides considerable support that the models were tracking visual processing of orientation. So it seems that people's implicit expectations can systematically bias motion processing

in early visual areas.

Given that the sound-orientation pairings were implicitly encoded, it's reasonable to ask where such expectations are encoded. However, one might think that the expectations are stored as inter-modular associations between auditory cortex and early visual cortex. If so, this need not be a problem for someone like Pylyshyn. The evidence suggests, however, that associative connections of this sort are not stored as direct links between otherwise encapsulated sensory systems. Rather, neuropsychological data suggests an essential role for medial temporal cortex, or parahippocampal cortex (or both), which are areas not generally thought to be part of the visual system (nor the auditory system), but rather form crucial components of the long-term memory system. Murray et al. (1993), for example, removed these regions from the brains of monkeys, and found a dramatic decrease in the ability to learn new statistical associations. Further, Schapiro et al. (2014) studied a human patient known as LSJ, who suffered complete bilateral loss of the hippocampal formation and surrounding medial temporal lobe. It was found that not only is LSJ incapable of forming new episodic memories (which was already known), but that she is also incapable of implicitly learning new statistical associations between events in her environment. So there's at least some evidence that points to the penetration of early vision by information stored outside the visual system itself.[6]

Again it might be possible to explain these results as effects of attention.

[6]This kind of implicit biasing of visual processing doesn't involve conscious beliefs or desires penetrating vision, of course. (Unlike Kok et al.'s (2014) study.) As a result, many supporters of visual modularity may not regard such cases as particularly interesting counterexamples. But given that much of cognition operates at an implicit level, however, I fail to see why this should make the effects any less important.

For expecting the overall pattern of motion on a given trial to be oriented at, say, 27.5° might lead one to attend more to dots whose motion is most consistent with that expectation, thereby according them greater weight in the processes that calculates the overall direction. But here, too, this cannot be an effect of mere spatial attention. Attending to a particular location within the display couldn't cause such specific effects. (How would the visual system "know" how to accord some dots greater weight on the basis of spatially-based attentional commands alone?) And again, even if the effect *is* an attentional one, it seems to require a semantic relation between the content of the expectation the specific changes to visual processing.

We've looked carefully at two studies that show expectations (both explicit and implicit) modulate visual processing. But recall from the beginning of the section, there are others. I chose to concentrate on Kok et al. (2013) and Kok et al. (2014) because these studies nicely pair behavioral and fMRI imaging techniques, which allow us to test the encapsulated hypothesis against the unencapsulated hypothesis. What we find is considerable evidence that vision is unencapsulated.

## 3.4 Visual Imagery and High-Level Image Content

To make the case that visual processing is modulated by the semantic content of high-level states, I have been discussing some subtle effects that are restricted to a very specific set of findings. It's important two work through these detailed experiments, but one might think that if high-level content affected perceptual processing on a regular basis, we'd find more evidence of cognitive penetration. I now

want to discuss two illuminating strands of research: research on visual imagery that, curiously, is often not taken as support of cognitive penetration; and research on semantic priming of ambiguous images.

### 3.4.1 Visual Imagery

When one visually imagines something—e.g., picturing a loved one's face or imagining oneself making a perfect free throw—concept-involving goals are used to construct a visual or quasi-visual representation. This much is uncontroversial. If I am to *intentionally* image making a free throw, the system responsible for the imagery must somehow be sensitive to that intention.

There is now ample evidence that the psychophysics of visual imagery exhibits properties we intuitively associated with visual representations (Kosslyn, 1994). In a now classic experiment supporting this hypothesis, participants were asked to determine whether the object depicted in one image was identical to, or a mirror image of, an object depicted in an adjacent image. In order to determine whether the objects were identical, participants were asked to mentally rotate the first image. Shepard and Metzler (1971) found that response times are a linear function of the amount of mental rotation required to transpose the first object into the same orientation as the second. If a participant has to rotate an image 180°, she will be slower to identify a match/mismatch than she would if she only had to rotate the image 90°, suggesting that the mental "imagery" is, indeed, an imagistic representation. There is also considerable evidence that visual imagery and vision share cortical mecha-

nisms (Mechelli et al., 2004; Reddy et al., 2010; Slotnick et al., 2012; Tong, 2013). So if concept-involving goals can drive visual imagery *and* visual imagery and vision share cortical real estate, it appears that top-down signals carrying semantic content can cause the visual system to activate semantically related visual representations.

Consider a study by Vetter et al. (2014). These experimenters trained pattern-classifers to identify what people were hearing or imagining from patterns of neural activity in the early visual system (V1, V2, and V3). The classifier was able to discriminate whether people were imagining a forest, or traffic, or a group of people talking. It seems that the high-level goal of forming such an image is capable of causing category-specific patterns of activity in early visual areas. We can conclude, then, that these areas are cognitive penetrable.

Albers et al. (2013) also used pattern classifiers to investigate neural activity in early vision in cases where people saw a grating, or held a representation of a perceived grating in working memory, or followed instructions to generate a mental image of a grating at a particular orientation. In each case the fMRI classifier was able to identify the orientation of the grating from patterns of neural activity in early visual cortex (either from V1, V2, and V3 collapsed together, or within each individually), and it did so with high degrees of reliability. Notably, when participants were visually imagining the gratings, patterns of activity in early visual cortex closely resembled the observed patterns in cases where people perceived a grating of the same orientation, suggesting that the same mechanisms are implicated in each. Moreover, this resemblance was greater for people who score higher on the Vividness of Visual Imagery Questionnaire.

It might be possible for a defender of encapsulation to respond that although vision and imagination depend on the same early-visual brain regions, the relevant neural populations within these regions are disjoint from one another. That is, it might be said that one set of neurons can be activated in a top-down manner for purposes of visual imagery, whereas a distinct set is involved in bottom-up visual processing, and the former cannot influence the latter.

This is of course possible in principle. And there is some theoretical plausibility to the suggestion. Even if one isn't committed to the visual system being encapsulated, on the whole visual processes seem somewhat insulated from cognitive processes involving long-term memory. Furthermore, visual imagery is remarkably different from normal visual processing in that it is entirely connected with high-level semantic memories that generate the imagery. So one might think that there is an imagery *system* that is distinct from the visual system.

We find some empirical support for the idea that the processes underlying imagery constitute a distinct system. Lee et al. (2012) find that while there is some overlap of neural activity when, say, imagining a desk lamp and perceiving a desk lamp, when one looks at activity across all visual areas of the brain one finds quite distinct activation patterns. They argue that their results show that visual imagery and perception are "distinct, suggesting that imagery is not just a weak form of perception" (Lee et al., 2012, p. 4071).

Note, however, that we really have two objections on the table. The first is that the neural substrates implicated in visual imagery are distinct from the neural substrates that support visual perception. The second is that visual imagery and

visual perception consist of distinct systems. This latter position is consistent with the two systems sharing neural resources.

Regarding the first objection, notice that it would require us to postulate that much of the functionality of the visual system is replicated in a separate imagery system. This is because the top-down patterns of activity in areas associated with visual processing are presumably at least partially responsible for the generation of mental images. (Recall that we know these areas correspond with reports of visual imagery from the Vetter and Newen, 2014; Albers et al., 2013, studies.) Thus, these mechanisms in early visual areas would need to be bound together and integrated into a coherent quasi-visual percept. Rather than posit machinery that parallels much of the functionality of normal visual processes, it is more plausible to assume that the same mechanisms are used for each. If memories are stored within the systems where they are initially processed, as many people assume, then memories of imagining a basketball bouncing should be stored separately from memories of seeing a basketball bounce. But in fact the two memories are readily confused, suggesting that they are realized in a similar manner.

With regards to the second objection, nothing in my argument requires that we equate normal visual perception with visual imagery. There are a number of different, but mutually compatible ways of individuating systems within the mind/brain, and how we individuate systems may depend on our theoretical interests (Craver, 2007). Let's suppose that there is a principled way of individuating the visual from the visual imagery system, where each share cortical resources to some degree. What would this demonstrate? Of interest to this discussion, it would demonstrate

that high-level conceptual content can systematically activate low-level visual areas. There would likely need to be a mechanism that regulates when high-level systems can activate visual processes, so that one is not routinely engaging his imagination or hallucinating. But if this is one's view, there's very little at dispute. I agree with the classical modularity theorist's principal observation that we cannot will ourselves to see anything we like. But unlike the classical modularity theorist, I don't think this demands that we posit encapsulation. Normal visual processing results from both bottom-up and top-down factors. What we see in the normal case is highly constrained by visual input. And we can't do much to change that input, except close our eyes or look away, etc. However, what the visual imagery literature clearly shows is that high-level cognitive states can alter, in semantically related ways, the content of visual processes. And this is all that I am claiming.

### 3.4.2 Moony Images

Many researchers use Mooney images to investigate the effects of semantic knowledge on perception. Recall that Mooney images (or two-tone images) are high contrast images made from normal photographs (by converting the color scheme to greyscale, maximizing high contrast, and minimizing low contrast). Without previous exposure to the original photo, a Mooney image generally looks like black splotches on a white background. Remarkably, even short exposures to the original image bring the primary figures into relief, suggesting that a memory of the image alters visual processing on subsequent viewing.

To investigate this possibility, Hsieh et al. (2010) used fMRI and pattern analysis to show that perceiving the meaning in a Mooney image alters processing in early visual cortex. The experiment involved three phases. First, participants were scanned while they viewed a series of Mooney images for the first time. Second, participants were scanned while they saw the full greyscale images from which the first set of Mooney images were constructed. Third, participants were scanned while they viewed the original set of Mooney images.

The pattern of activity in early visual cortex during the final phase (when the Mooney image was meaningful) was more similar to activity when perceiving the corresponding greyscale image than it was to the pattern of activity when perceiving the Mooney image in the first phase. Since the external stimuli in the first and final phases were identical, bottom-up processing should likewise have been the same. So the shift in the pattern of activity in the final phase can only result from the top-down influence of participants' knowledge of the meaning of the stimulus.

Presenting work from his lab, Christoph Teufel finds that perceiving Mooney images as meaningful stimuli alters low-level edge detection.[7] In this experiment, participants were asked to identify the orientation of a small "edge" placed on Mooney images where a meaningful contour would be located in the full greyscale version, before and after exposure to the full greyscale versions of the image. The "edges" (small Gaussian contrast patches) were either aligned with, or orthogonal to, the unseen contour. Varying degrees of noise were randomly assigned to the patches to alter the difficulty of identifying the orientation. Teufel and colleagues

[7]From work presented at The Cognitive Penetration Workshop, Bergen, June 2015.

hypothesized that contour detection would be better *after* exposure to the full image when the contour is *congruent* with the unseen, but "meaningful" contour. This is, in fact, what they found, suggesting that a memory of the image facilitated contour processing for contours that were congruent with the unseen contour.

Teufel notes that he and his colleagues controlled for focal and feature-based attention by holding attention constant across the "before" and "after" conditions.[8] To control for focal attention, when viewing the Mooney image, a fixation square directed visual attention to the same patch location in both the before and after conditions. (The timing of the fixation square's appearance also predicted the appearance of the contour patch.) Given that focal attention was directed to the same location in both conditions, it's hard to see how attention could explain the difference in discrimination performance. To control for feature-based attention, they indicated which two possible orientations the participant might see. Of course, they did this for both before and after conditions. So if one thought that the participants' predictions about what they were about to see was engaging feature-based attention, then one would expect the same results in both conditions. But, again, this is not what they found.[9]

Neri (2014) ran a similar study where participants had to identify the orientation of a small contour patch, though this time the patches were embedded in natural scenes. They sought to explain why orientation discrimination for contours

[8]C. Teufel, personal communication, November 16, 2016.

[9]It's always possible that there could be some other form of attention at work in these kinds of results. But one problem with the "it's just attention" objection at this point is that it's quite unclear what the alternative explanation amounts to. Simply saying "it could be attention" without a proposed mechanism threatens to broaden the notion of attention to the point that it no longer predictive anything in particular.

that are congruent improved with contextual features in the scene (e.g., the continuation of a contour). One possibility is that contextual features are processed in a bottom-up fashion, which in turn modulates processing in different parts of the scene (similar to the model discussed in relation to Li et al.'s (2004)). Alternatively, the visual system might generate a rapid high-level "gist" representation, which is, in turn, fed down to modulate local contrast processing in early visual areas.

Inverting a natural scene image disrupts gist processing, making it more difficult to recognize whether one is looking at, for example, a desert or a forest. Neri hypothesized he could disrupt high-level semantic gist encoding by inverting the images, and this would then neutralize the top-down effects on lower visual processing. He found that the data supported his hypothesis: participants are better at identifying the orientation of embedded "edges" when the edge align with a contours in right-side-up images, as opposed to up-side-down images.

One might argue that both the Teufel and Neri studies merely show that an imagistic memory of the scene is doing the work in these experiments. A structural-spatial description, for example, of a women kissing a horse on the nose, could be stored within the visual system, such that it primes early visual process for the contours at a particular spatial location in subsequent Mooney images.

This is a fair point. But even if the visual system is responsible for storing visual memories, it's not clear that this counts as a vindication of classical modularity. First, such a view would entail that the visual system encodes long-term imagistic memories, and these memories can be recalled by ambiguous contexts. This would involve a substantial departure from the modularist's understanding of

visual function.[10] Second, we know that perception of meaning in Mooney figures can be be secured not only by showing the original greyscale picture, but also by conceptual priming of various sorts (such as being told that the hidden-Dalmatian figure contains an image of a dog, or by hearing a dog barking) (Long and Toppino, 2004). And there are also demonstrations that conceptual priming can bias the perception of ambiguous figures such as the duck–rabbit (Balcetis and Dale, 2007). So it would seem that even if we expanded the traditional role of the visual system, we find evidence of top-down conceptual effects on image disambiguation.

## 3.5 A Revolution?

I have been arguing that the visual system is not encapsulated. Goals, expectations, and conceptual memories can and do effect visual processing in a semantically-driven way. Some argue that such a position undermines any coherent distinction between perception and cognition. For example, Firestone and Scholl (2015) write that "the extent to which what and how we see is functionally independent from what and how we think, know, desire, act., etc." bears on "whether there is a salient 'joint' between perception and cognition" (p. 8). That is, Firestone and Scholl think that one can draw the traditional distinction between perception and cognition only if perception is encapsulated from cognition. They also note that bottom-up models have been remarkably successful in modeling many standard aspects of visual

[10] I am by and large friendly to this sort of view (see Chapter 7). If this alternative explanation of the Mooney image data is correct, then it provides support for my view that we visually represent high-level properties. Either way, this sort of evidence supports the general picture of vision that I am advancing in this dissertation.

processing, and suggest that if cognitive penetration were to occur, then traditional models of perceptual processing are radically misguided or incomplete. They seem to think, then, that if cognition "can affect what we see... then a genuine revolution in our understanding of perception is in order" (p. 5).

While certain views might entail the abolishment of the perception–cognition distinction, in particular the radically interactionalist views of Clark (2013) and Hohwy (2014), I see no reason to think that the viability of a distinction between perception and cognition depends on perception being encapsulated from top-down information. As I noted in the previous chapter, there are weaker version of modularity that don't require encapsulation. One can characterize the visual system as the set of brain-mechanisms specialized for the analysis of signals originating from the retina. The computations these mechanisms perform are geared towards making sense out of the light array landing on the retina. We may not know precisely how to identify these mechanisms or how they perform their computations, but one can accept that the visual system consists of a proprietary set of mechanisms while denying that it takes only bottom-up input. For example, the existence of cross-modal effects need in no way undermine the distinction between audition and vision. Hence I see no reason to think that the existence of top-down effects should undermine the distinction between vision and higher-level cognitive systems, either.

Of course, if there were no way to identify some set of mechanisms as proprietary to the visual system—i.e., if there's no way to draw a distinction between perception and other cognitive systems—then one might be justified in denying the traditional distinction between perception and cognition. But I see no reason for

such skepticism. In fact, holding fixed (or abstracting away from) top-down effects provides one effective way of individuating perceptual systems. Having established a relatively plausible model of bottom-up visual processing, one can thereafter look at how top-down factors modulate that processing.

Indeed, this appears to underlie the methodology employed by Kok et al. (2013) discussed above. Recall that they used fMRI data for input into a "bottom-up" model of motion-direction processing. Kok and colleagues assume that such a model will predict behavioral responses in the absence of a tone because the system (comprising at least V1, V2, and V3) is specialized for processing visual inputs, and will do so relying on bottom-up information alone in the absence of top-down modulation (which is what they found). In the presence of a tone, however, the model continues to match the behavioral response, but no longer tracks the stimulus orientation. Thus, they infer that there must be some sort of exogenous signal that alters the manner in which information is processed within the visual system.

In short, rather than obviating any distinction between perceptual and cognitive systems, this model seems to presuppose such a distinction, all the while allowing for top-down information about the statistical regularities in the environment to bias its computations. In fact, it is in virtue of stable bottom-up models that one can begin to understand how top-down effects modulate visual processing. This example also nicely demonstrates the compatibility of traditional bottom-up models with a more interactionalist view of visual processing. Thus, one should not be skeptical of the evidence presented in this chapter on the basis that it would require a revolution in the vision sciences.

## 3.6 Conclusion

The body of literature reviewed in this chapter supports the idea that the outputs of visual processing are under what we might think of as "soft" cognitive control. Expectations and goals influence the values of low-level visual parameters, how we parse a visual scene, and which features or spatial relations are singled out for preferential processing.

Expectations and goals don't dominate visual processing or determine what we see. We cannot will ourselves see what is not there. False expectations will not cause us to hallucinate (at least not in the non-pathological case). Nevertheless, this form of soft control does mean that we are not mere passive recipients of perceptual information. The sorts of things we think about, and the sorts of habits we form influence how we see the world. As such, we are (at least in part) epistemically and morally responsible for how we see the world. To what extent we are responsible? Under what conditions is it appropriate to hold someone response? These are difficult questions, but ones worth exploring. But attempting to address them in a systematic way calls for more attention than I can devote here.

# Chapter 4: Framing the Perceptual Content Debate

The previous two chapters dealt with one strand of the traditional dichotomy between perception and cognition, namely, that perception is not isolated from cognitive influences. But those chapters merely investigate the effects of cognition on standard visual processes. Everything I have said thus far is consistent with the visual system being a more or less an unintelligent subsystem of the mind.

This chapter and the following three address the scope of visual content. The central issue in this debate, at least for our purposes, concerns whether the empirical evidence licenses us to attribute high-level content to the visual system. And this issue, I shall argue, speaks to the broader question of whether vision is a genuine epistemic system. In the remainder of this chapter, I shall chart some of the logical landscape of the debate, and evaluate two arguments for why we should think that visual content is low-level.

## 4.1 The Debate

The dispute about the representational power of visual perception largely boils down to whether perception can represent so-called "high-level" properties. Theorists generally agree that vision at least represents the geometric and chromatic

properties of the immediate environment; however, philosophers and psychologists remain divided as to whether vision can represent high-level properties, such as *being an individual* (e.g., *being Barak Obama*), kindhood (e.g., *being a pine tree*), causal relations (A *moves* B), animacy (*having basic goals*), social properties (e.g., *being happy* or *having intentions*), or "ecological" properties (e.g., *being dangerous*, *edible*).

As mentioned in Chapter 1, the default view in psychology and empirically-minded philosophy seems to be that vision represents the geometry and chromatic aspects of our immediate environment, and that's about it. The reason, I think, has to do with the traditional assumption that there exists a close relationship between the laws of optics and perceptual content. Consider the following gloss one might have on the role of the visual system.

The light hitting our eyes contains structure, and it's the job of the visual system to determine the distal cause of that structure—namely, the geometric and chromatic structure of the immediate environment. Of course, the structure within the light array does not uniquely specify the layout of our environment at any given moment. For example, a dinner plate viewed at an oblique angle will produce an oval two-dimensional retinal image, which is consistent with an infinitely large set of possible causes, such as an oval plate viewed from straight on. The "retinal image" is ambiguous in many other ways as well: identical changes in contrast (or chromaticity) can be caused by shadows or changes in surface reflectance properties; small objects viewed up close will subtend the same angle on the retina as large objects far away; and movement across the retina can arise from a moving object or

from moving eyes. In short, the proximal stimulus radically underdetermines what we see. This is known as the "inverse optics problem."

Vision scientists standardly see their job as articulating the principles by which the visual system constructs a structural description of the distal causes of the proximal stimulus, and this project is apparently motivated by the inverse optics problem. Consider the following introductory passage from a review of research on computational vision science:

> This paper describes an alternative that has been developing over the last 20 years within the computer vision community. It treats perceptual interpretation as a solution of an inverse problem that depends critically on the operations of *a priori* constraints. (Pizlo, 2001, p. 3145)

Although nothing about the inverse optics problems demands that visual processes be restricted to low-level content—indeed, high-level properties, too, are radically underdetermined by the proximal stimulus—it's noteworthy that Pizlo limits his discussion to standard low-level visual features (binocular depth perception, motion perception, color and lightness perception, figure-ground segmentation, shape perception).

*If* one thinks that a complete solution to the inverse optics problem only involves articulating the principles necessary to fix a structural description of the distal causes of the retinal image, then it's hard to see how a such an empirical theory will have the theoretical resources to explain categorical or abstract content. Whether one is looking at a *genuine* Picasso or a *very good duplicate* is irrelevant

for such purposes. To put the issue differently: a genuine Picasso or a clever forgery will produce identical retinal projections (holding constant eye movements, viewing distance, etc.). Thus, we can swap out the original for a good forgery and it won't make a difference as far as the initial sensory impingements are concerned. So if the point of the inverse optics problem is to trace the etiology of the proximal stimulus—i.e., the features of the environment that make a difference to the retinal image—then the property *pained by Picasso* is not part of that etiology. And if perceptual content is restricted to properties in this etiology, then the prospects are dim for finding high-level visual content.

Clearly something is correct about this way of modeling visual perception. We know the laws of optics place considerable constraints upon visual perception, and therefore this makes a good place to begin one's inquiry. But why think that visual content is restricted to properties that are the causal origins of the *proximal* stimulus? The retinal image contains valuable structure, to be sure. But extracting the structure from the retinal image is not an end in itself. In order for the proximal stimulus to be of any use to an animal, it must provide clues about what is in the environment. So it's unclear why visual content should be restricted to the geometric and chromatic distal causes causes of the proximal stimulus.

At any rate, I argue in Chapter 7 that there is now good reason to broaden our theoretical commitments concerning perceptual content. The aim of this chapter, however, is to clarify the nature of the dispute. In what follows, address a variety of preliminary issues concerning the high–low debate, lay out the main case for thinking that perception is restricted to low-level content, and argue that this line

of thinking is not well motivated.

## 4.2 What is Perception?

The high–low question asks: what is the content of *perception*—or in this case, visual perception? But this question isn't yet clear enough, as different theoriests different things in mind when attempting to characterize "perception." For instance, one might be concerned with describing what someone is visually aware of. A prosecutor might want to know whether a witness saw (was aware of) the defendant at the bar on the night of the murder. Understood in this broad sense, conscious visual experience seems to be rich with semantic content. When I look out at the world, I experience (or seem to experience, at any rate) individual people, cars, trees, dogs, and so on. But what I'm visually aware of at any given moment (in a broad sense), may be an amalgam of beliefs, emotions, and perceptual states. Such an amalgam may not be amenable to serious scientific study, in the way that being in love may not be amenable to serious scientific study.[1]

In contrast to conscious visual experience, some theorists are interested in the "phenomenal" aspects (or character) of conscious visual experience—the "what it's likeness" of seeing a green leaf or a round ball. Essentially, the phenomenal character is the felt aspect of sensory perception, for example, the experience of a ripe

[1]There is also the worry about how we use ordinary speech to describe what we perceive. It might be true that I see Hank Aaron's 755th home run ball—i.e., it's true, de re, that I see that particular object—but most theorists would deny that I perceptually represent this fact. On the standard view, I perceptually represent the ball's geometrical and chromatic properties and form a belief about its significance. But I needn't perceptually represent the fact that the ball is a Hank Aaron home run ball. Indeed, I may not know the purpose for which the ball is used, or that it has any significance.

tomato's redness, or the smell of fresh coffee. Phenomenal *content* is defined as "that component of a state's representational content which supervenes on its phenomenal character" (Bayne, 2009, p. 386-87). The idea here is that there can be no change in the phenomenal content without a change in the phenomenal character. Phenomenal content, then, is generally understood to be restricted in what it represents. A red patch in my visual experience might represent the redness of a flowerpot, but it doesn't represent the flowerpot, as such.

The concepts of phenomenal content and phenomenal character (and consciousness more generally) are fraught with controversy. No one is quite sure how to explain the relationship between content and conscious experience. Indeed, it's controversial whether phenomenal character has content. An aggravating factor in all this is that these concepts are motivated by appeal to introspection and intuition, which makes resolving disputes rather difficult. I see very little progress to be made by hacking away at this Gordian knot.

A more tractable subject of inquiry is to ask what the visual system does, how it operates, and how it encodes information from from the light array. Answers to these questions can then be used to address the broader question of what the visual system represents. Of course, vision science has only a fragmentary understanding of the operations of the visual system. And so the conclusions we draw will be modest, and at times, less secure than we might otherwise desire. But theory construction requires that we start somewhere. My only recommendation is that we look to the vision sciences to inform the way we frame our questions and how we answer them. Such an approach has a distinctive advantage over a priori and introspection-based

theorizing: the vision sciences offer a number of rich theoretical frameworks across which one can look for converging lines of evidence.

In the following chapters I will be reviewing a variety of findings from the vision sciences, and explicating their underlying theoretical commitments. Henceforth, the primary theoretical question is: what does the visual system represent?[2]

## 4.3 Representation

The high–low debate assumes that vision is representational. This is a relatively uncontroversial assumption in the vision sciences and the philosophy of perception. So while I won't defend this thesis in depth, I ought to provide some motivation for the thesis and say a little about my specific commitments.

To claim that vision is representational is to claim that a state of the visual system is *about* some object, property, or event.[3] In philosophical circles, the aboutness relation is spelled out in terms of the semantic properties of the state, such as its truth or correctness/veridicality conditions. That is, a state represents (or is about) $x$ by virtue of possessing the appropriate semantic properties.

This rather abstract description of a representational state is all well and good, but it's not particularly helpful when attempting to empirically identify something as a representation. If a system or state possesses semantic properties, we will have to infer them on the basis of empirical evidence. So if we are to bring empirical

[2]On occasion, I will engage claims concerning "phenomenal character/content" since this is one of main ways that philosophers address the issue of visual content.

[3]Note that an object, property, or event needn't be instantiated in order for a mental state to represent it.

evidence to bear on this issue, what whould the evidence look like? I suggest the following three criteria.

*Distality*: One reason to think a system is representational is if it tracks (or is apt to track) distal objects, as opposed to mere sensory registrations. If state of the visual system is best described as processing information concerning properties of external objects, this suggests that the the state is about that object.

We have good reason to think that the visual system is geared towards discriminating and tracking distal objects. Vision scientists aim to model (among many other things) shape perception, figure-ground segmentation, object tracking, shadow-contour disambiguation, motion perception, and depth processing. It's hard to see these as sensible research areas if the visual system is not in the business of tracking distal objects.

*Misrepresentation*: Representation implies the possibility of error. So evidence that the visual system *mis*attributes properties to distal objects gives us reason to think that it is representational. Illusions and other sorts of psychophysical curiosities provide examples of misattribution.

Looking at the Müller–Lyer figure, the two lines *appear* to be of different lengths. That is, the world appears to be one way, when in fact it isn't. Perceptual adaptation aftereffects provide another instance of inaccurate perception. For example, if one fixates on a tilted line (say, 15°to the left) for 30 seconds, subsequently viewed vertical lines will appear tilted slightly to the right. Again, the world appears one way, when it is another. So cases of illusion and perceptual adaptation suggest that we can visually *misperceive* the world, and the fact that our visual systems can

get things wrong suggests that we visually represent the world.

*De-coupleability*: Typically, representations are posited to explain flexible, future-orientated behavior, behavior that cannot be explained in terms of dispositions or brute stimulus-response profiles. Pretty much any sort of complex planning is thought to involve representations. Consider, for instance, Santino, a male chimp in the Furuvik Zoo in Sweden. Santino is the dominant male in his troupe. He shows his dominance to other chimps by throwing projectiles in their direction. The goal, it seems, is not to cause harm, but to cause the other animals to move, thereby demonstrating his dominance. Zoo staff members were understandably concerned when Santino attempted to prove his dominance to some of the human visitors to the zoo. So prior to visiting hours, staff would collect any potential projectiles. It didn't take long, however, for Santino to stymie the zoo keep's efforts. Observers began noting that prior to visiting hours, Santino would collect and subsequently conceal stones and various other potential projectiles. It was clear that Santino was intending to use these objects to display his dominance (Osvath and Karvonen, 2012).

It's precisely this sort of flexible, future-oriented behavior that calls for representational content, psychological structures that stand in for various environmental features (the zoo keepers, the other chimps, the human visitors, rocks, etc.), which can then be utilized in a flexible manner in reasoning, planning, and action. Santino has the belief that there's a rock under a pile of hay, because he believes he put it there early this morning. And he can access this belief when his visitors arrive.

This example clearly motivates positing non-perceptual mental states, such

as beliefs and goals. But it's unclear the extent to which perceptual states are flexibly deployed in the way that, say, concepts are flexibly deployed. If perceptual states figure in, say, action planning, they don't figure in the way that concepts or memories figure in these sorts of processes. Calling upon a current perceptual state is of no help if the object of one's action is not in sight.[4] The perceptual states that guide Santino actions when he hides his projectiles are not the states he enlists when carrying out his plan. If this is right, then, we can all agree that Santino has beliefs and desires, but this doesn't demonstrate that Santino perceptually represents his environment.

There is, perhaps, a more fundamental feature that perceptual and cognitive representations share. Orlandi (2014) suggests that de-coupleability is a central feature of representation: "A state is de-coupleable, at a minimum, when the state does not require constant causal contact with what it is about" (p. 122). The idea is that the representation of $X$ has to allow for the absence of $X$. Beliefs are clearly de-coupled in this sense. We can think about things past, future, and entirely remote: "Rome was once the center of a great empire, and may host the 2024 Summer Olympics."

Plausibly, perceptual states can be decoupled as well. We represent the entire cat even though we only see bits of it through the picket fence. We represent the entire apple, even though only one side is visible. Alternatively, the percepts generated by the Kanizsa Triangle display are entirely de-coupled from what they are

[4] I'm appealing, here, to a traditional understanding of perceptual states, where the primary role is to "see the world." Note, however, that we have some reason to think that perceptual states are involved in prospection and planning (Moulton et al., 2009). If this is right, we have further reason to think that perceptual systems are representational.

about. When viewing this display, people normally see a triangle partially bounded by illusory contours: the "packmen" at the vertices are "connected by" non-existent contours. Since there are no contours that cause this aspect of our visual experience, there can be no causal relation. This is de-coupleability with bells on!

Of course, the visual representation is dependent on the incoming stimulus. But the point of the de-coupleability requirement is not that there can be no cause of a mental representation (all events have a cause); rather, it's that the *content* of the representation is not stimulus bound.

Even if one thinks these cases can be explained without appealing to representation, de-coupleablity strikes me as an appropriate necessary condition on representation. One reason is that de-coupleablity is closely connected to misrepresentation. In order for a state to misrepresent, it must be decoupled from the world. If I mistake a horse for a cow on a dark night, there can be no causal connection between cows and my thought about cows, since, by hypothesis, there are no cows in my environment. In Chapter 7, I argue that we represent high-level properties by virtue of possessing schemata that organize the incoming sensory information. I claim that these schemata generate de-coupled perceptual representations.

So finding evidence of distality, error, and de-coupleability will provide us with reason to think we are dealing with a representational system. And, as I have noted, visual processes seems to exhibit these properties. Before moving on to the next issue, I should say something about the format of visual representation.

The computational/representational theory of mind (CRTM) is the most explicitly formulated theory of belief–desire psychology (Fodor, 1975). According to

CRTM, complex behavior is the result of computations over structured representations or attitudes (i.e., thoughts and desires), which are composed out of simpler representational constituents. Intentionality, on this sort of view, is a fundamental posit. Beliefs, desires, and their constituents are intentional through and through. Some think that a similar explanatory approach can be applied to explaining visual processing (Clark, 2000; Marr, 1982; Rey, 1998).[5] Very roughly, the picture goes like this.

The visual system utilizes information transduced at the retina to predicate visual features and properties to locations in represented space. For example, retinotopically encoded contrast gradients in the light array provide clues to the location of surface edges. Integrating other spatial, shading, and luminance cues, the system predicates EDGE to a particular region of visual space. Once a bounded region has been obtained, SURFACE is predicated to that region in represented space. Colors, textures, etc. follow suit.

Whether or not this is a successful explanatory strategy will depend on how the details are spelled out. However, for my purposes, I want to remain much more minimally committed to representationalism. That is, I am committed to vision being representational, but I remain neutral on whether CRTM or some variety of connectionism (or some mixture of the two) best accounts for the representational capacities of visual processing.

[5]Of course, other theorists think that vision is computational (e.g., Burge et al., 2005; Fodor, 1983; Pinker, 1997; Pylyshyn, 1980) but they seldom spell out what such a view would look like for visual processing.

## 4.4 High- and Low-Level Content and Properties

As I discussed earlier in this chapter, the main debate concerns whether we visually represent high-level properties. Following Bayne (2009) we can characterize the debate as follows. The conservative view holds that perceptual content—and in particular visual content—is restricted to the low-level properties. The liberal view holds that, in addition to the standard collection of low-level properties, visual perception represents at least some high-level aspects of our environments.

At this point, one might legitimately wonder what the difference is between high- and low-level properties. Logue (2013) suggests that the distinction is just a "shorthand way of referring to the difference between properties pretty much everyone agrees we can visual experience, on the one hand, and properties that not everyone agrees we can visual experience, on the other" (p. 2). I think that Logue is more or less correct about this issue, but with one caveat. It's not merely that we find agreement about what vision can minimally represent; we think representing high-level properties requires concepts or concept-like entities. Nearly everyone agrees that to see something *as* a car, requires the application of a concept. But if we visually represent high-level properties, then our visual systems must trade in concepts or concept-like entities. This claim is controversial, and is what the high-low debate boils down to.

## 4.5 Content Conservatism

As Siegel (2010) notes, the traditional view amongst philosophers is that visual content is restricted to low-level properties. I call this view content conservatism. While I don't know the extent to which theorists are committed to content conservatism, it's clear that prominent theorists hold the view.

In her discussion concerning the history of the Rationalist–Empiricist debate, Carey (2009) characterizes what both sides of this historical debate understood as "sensory primitives":

> Sensory representations may be roughly characterized as those representations that are the output of the sense organs. They are what psychologists call proximal representations—those representations that maintain the point of view of the pattern of stimulation on sense organs. (p. 29-30)

Carey goes on to embraces this view. But note that this view is quite restrictive, as it rules out any sort of perceptual constancy mechanisms that account for *perceived* lightness, size, and depth—aspect that are otherwise understood to be paradigmatic perceptual processes. Carey goes on to say that "[O]bject representation, like depth representations, are clearly non-sensory, for they represent distal entities" (p. 34). Early Spelke (1988), too, holds that object perception must be conceptual (and therefore cognitive) because apprehending the same object despite transformations in appearance across time requires an understanding that objects are "*cohesive,*

*bounded, substantial, and spatiotemporally continuous*" and these abstract properties are not the sort that can be picked up by the sensory organs (p. 226).

Burge (2010) is also skeptical that we represent much beyond the standard low-level properties are represented in perception:

> Purely perceptual representational contents represent only attributes that an animal can discriminate as a result of processes that begin with sensory states that are sensitive to a specific causal medium—light, sound, contact, and so on. Most visual perceptual systems form representations of a small number of environmental attributes—integrated body, shape, spatial relations, motion, texture, brightness, color, and perhaps functional properties like food, danger, shelter. (p. 101)

According to Burge, perceptual content is constrained by the aspects of the environment that impinge upon our sensory organs, echoing my earlier remarks about the supposed close relationship between the laws of optics and perceptual content. Note that, contra Carey, Burge (2011) thinks that "[T]here is substantial evidence that perceptual body representations occur in the visual systems of many mammals and some birds" (p. 125).[6]

Carey and Burge are far from alone on this issue. Prinz (2006), argues that, as far as phenomenal experience goes, vision represents "colors, shapes, textures, and sounds from particular vantage points" (p. 451). And Dretske (2003; 2015), McGinn

[6]In response to Burge's criticism, Carey concedes "that perceptual representations go beyond sensory ones in these ways, and that therefore these facts do not rule out that object representations in infancy are perceptual" (p. 155). But even with this concession, she is reluctant to admit much else in our perceptual representational schemes.

(1982), Tye (1995) among others have claimed that vision is restricted to low-level features. Thus, we should take perceptual content conservatism as a serious view.

## 4.6 Two Arguments for Content Conservatism

We can identify two different argument strategies for holding a conservative view about perceptual content. The first argument is a species of substitution argument. We see (or perceptually represent) only those properties or features that make a difference to our experience. For example, if I am looking at an apple, I don't visually represent it *as an apple*, because something other than a genuine apple (e.g., a wax apple) could look the same. The second argument appeals to information processing considerations to constrain what can be represented in perception.

### 4.6.1 Substitution Arguments

When I think of substitution arguments I think of work by two Swiss artists, Peter Fischli and David Weiss. The two created rather elaborate workshop scenes out of painted polyurethane foam. They sculpted and painted blocks of foam to create objects that appeared like old paint cans, cigarette packages, brushes, blocks of wood, steel wrenches, furniture, etc. The work is utterly convincing to the naked eye, and offers a clear example of how surface appearances can be preserved while the underlying reality altered. Fischli and Weiss have substituted out the properties of *being a paint can* or *being a made of wood* without altering how the objects appear. Since these sorts of properties can be substituted out, many take this as

reason to think that we don't perceptually represent these properties. Those who take advance this line of thinking subscribe to the following principle:

> SUB: Holding all else equal, an organism perceptually represents only those properties of the environment the removal of which make a difference to how things appear.

Substitution considerations have a long history in philosophy. We find something like SUB at work in Descartes' (1984) *Second Meditation*:

> But then if I look out of the window and see men crossing the square, as I just happen to have done, I normally say that I see the men themselves, just as I say that I see the wax. Yet do I see any more than hats and coats which could conceal automatons? I judge that they are men. And so something which I thought I was seeing with my eyes in in fact grasped solely by the faculty of judgment which is in my mind. (p.21; AT 32)

Descartes seems to be saying here that we cannot legitimately claim that we perceive men, because we could remove the men but preserve the appearances.

Descartes, of course, had a epistemological agenda. Because he envisioned a foundationalist epistemology (inspired by Euclid's "geometric method"), he was in search of beliefs about which he could be certain—beliefs that would form the foundation of his system of knowledge. In order to identify those beliefs about which he could be certain, Descartes subjected his beliefs to radical skepticism. Can we be certain that the "men crossing the square" are genuine men and not

automatons? No: one's visual experience in this case are entirely consistent with a radically different underlying reality.

This epistemological project was popular in the first half of the 20th century philosophy. In *The Problems of Philosophy*, Russell begins with a Cartesian epistemological concern: "Is there any knowledge of the world which is so certain that no reasonable man could doubt it?" (Russell, 1912, p.1). He then proceeds to sketch a theory of perception, whereby the only thing perceived (i.e., known with certainty) are "sense data." All other aspects of the world, according to Russell, must be inferred on the basis of sense data.

We continue to find SUB applied in the contemporary high-low debate, as well. Brogaard (2013) appeals to SUB in arguing for why we are not "phenomenally conscious" of high-level properties. In its barest form, the argument goes like this: it's possible for two people to have identical visual experiences when looking at a real tiger and the other is looking at a fake tiger. If these agents represented high-level properties, then they would represent different high-level properties. But this means that phenomenal content does not supervene on phenomenal character (i.e., no change in phenomenal content without a change in phenomenal character). But since the supervenience thesis is true, we can't represent high-level properties.[7]

In a rather illuminating passage, Tye (1995) also endorses a very similar principle to SUB:

It seems plausible to suppose that the property of being a tiger is not

[7]For what it's worth, I don't find this argument to be compelling. But I shall address the general strategy of substitution arguments below.

> itself a feature represented by the outputs of the sensory modules associated with vision. Our sensory states do not track this feature. There might conceivably be creatures other than tigers that look to us phenomenally just like tigers. (p. 141)

What is remarkable about this passage is that Tye takes substitutability as not only a constraint on what we phenomenally experience, but also a constraint on the outputs of sensory processing. This suggests that substitution considerations drive intuitions about the scope of information processed in visual systems.

A general concern with SUB is that it seems to be rooted in a very specific epistemological project. But it's not at all obvious that Descartes' epistemological goals are well suited for modern theorizing about perceptual representation. Why should we think indubitability is the right criteria for identifying which features of the environment are perceptually represented? I may be entirely uncertain as to whether it will rain tomorrow, but I can entertain this possibility, and therefore represent it. Thus, questions about indubitability look to be orthogonal to questions about representation.

This undermines the motivation for SUB, but it doesn't show that SUB is false. As I shall now argue, SUB is too strong, in that it entails that perceptually represent very little of our environments, if anything at all. Let's consider an uncontroversial case of a perceivable property: shape.

Suppose you are looking at a square. We can then ask: is it possible to substitute squareness with some other property, all the while preserving the original

appearance? Answer: yes. A quadrangle viewed from an oblique angle will produce the same appearance. And it seems that we could do this for nearly any possible property. Because perception is perspectival, any shape can be substituted for another. Because lighting affects color appearances, one can perceive green things as red under the right lighting conditions, and so on. In fact, virtual reality technology is in the business of delivering to us illusory experiences. We can substitute out shapes, colors, tables, chairs for mere appearances of these properties or things.

If SUB is true, then we don't represent any of these properties, as they can all be substituted out with other properties without disturbing appearances. This is a striking conclusion. SUB, as Brogaard and Tye are applying it, is supposed to inform us about which properties are perceptually represented. It now looks like SUB entails that even the standard set of low-level properties aren't perceptually represented. This result, of course, doesn't show that SUB is false; Brogaard could double down on her commitment to SUB by taking the position that we only represent phenomenal appearances. Phenomenal appearances, here, would be properties that objects possess by virtue of how they appear in phenomenal visual consciousness. Even if the rectangle isn't a square, it has the property of *appearing square-ish*. And even if the region of the wall isn't red, it has the property of *appearing redish*. While we don't represent being a square or being red, we do represent phenomenal appearances.

This response preserves the consistency of SUB, but it does so at rather large cost. First, it's unclear if SUB actually rules out high-level properties. For instance, if we allow properties like *appearing square-ish*, then it seems that we can allow

properties like *appearing dog-ish* or *appearing pine tree-ish*. This is because these properties satisfy SUB: If a robotic dog looks like a dog, then it has the property of *appearing dog-ish*. *Appearing dog-ish* is preserved in this case, and so one could argue the visual system represents this (admittedly) rather strange property.

Second, it's not obvious in what sense phenomenal appearances are representational. For a state to count as representational, it must have correctness conditions. But SUB pretty much guarantees that a visual state is veridical, as objects have phemonemal appearance properties by virtue of how those objects appear. Now, trivially satisfied correctness conditions are nonetheless correctness conditions (Neander, 1995). But consider our earlier discussion of when it's appropriate to posit representations. I noted Orlandi's suggestion that in order for a state to be representational, it must be possible to de-couple the mental state from its normal causes. If we characterize content in terms of "appearance properties" then it's impossible to break that causal connection. Anything that appears, for example, square-ish (relative to an observer) will generate the corresponding perceptual content; and, things that don't appear square-ish won't. If we take de-coupleability as a necessary condition for representation, then phenomenal appearances are not representational.

I suggested de-coupleability as a defeasible empirical evidence for representation, so I don't think this argument is decisive. That is, when we find de-coupleability, we have reason to think the state is representational. We may still think a state is representational if it fails to de-couple from its normal causes, but only if we have independent reason to think the state is representational. This leads

to a further problem with phenomenal appearance properties, namely, that it's unclear whether we have independent reason to think the visual system represents these odd properties. They seem entirely ad hoc and unexplanatory. For instance, it's unclear what role a representation of *appearing dog-ish* is supposed to play in one's overall mental economy. One can easily see the motivation for positing representations about dogs. We think about dogs, we see dogs, paint dogs, and plan for the care of dogs. But it's not so obvious why we would posit these appearance properties. Representations of appearance properties were introduced as a way to preserve SUB. But if this is the only reason for positing them, then the move is ad hoc.

Presumably, what motivates SUB for contemporary theorists is that it supposedly captures intuitions about what counts as perceptual content, namely, the standard set of low-level properties. But once we recognize that SUB is much stronger than originally thought, it loses much of its intuitive appeal. Hence, we should abandon SUB as a constraint on a theory of perceptual content.

This might strike some as extremely counterintuitive, and quite possibly false. One might reply to my argument in the following way: surely substitution considerations are relevant to perceptual content. For if an experimenter finds that changing some aspect of a stimulus makes no difference to participants' reports, then we have good reason to conclude that the change was not seen, and therefore not visually represented. If participants cannot discriminate between two physically different color chips, the conclusion ought to be that they don't represent a difference between the two chips.

This is the correct response to someone who thinks that substitution considerations are *never* relevant for specifying perceptual content. But this is not what I am claiming. SUB makes the universal claim that any undetectable substitution instance tells us that the substituted property is not perceptually represented. I'm denying this strong claim.

Furthermore, notice that the color chip example doesn't ask whether a certain color is perceptually or non-perceptually represented. The experimenters have good independent reason to think that color is perceptually represented. It's now a question of whether normally sighted perceptual capacities can make fine enough discriminations to tell the two chips apart. The above theorists (Descartes, Russell, Brogaard, and Tye) might think that their employment of SUB is a natural extension of this methodology. They want to know if manipulating some environmental feature will make a difference to the visual system. Will a (visually identical) counterfeit tomato have the same effect on our visual system as a genuine tomato? This is a completely legitimate question to ask, and the correct answer, I presume, is that the two objects will have identical effects. But it doesn't follow that we only represent the low-level properties of the tomato, at least not without some auxiliary premises.

One such auxiliary premise might be the following: the only thing that the counterfeit and genuine tomato have in common is their low-level properties; hence, we must only represent the low-level properties. But recall our earlier discussion of SUB. A square and a quadrangle (viewed at an oblique angle) can have the same effects on our visual system. This case highlights the fact that we have a choice

between saying that we only represent the common property that holds across both cases or we *misrepresent* a rectangle as a square. We are faced with the very same choice when we think about the tomato example. Do we merely represent the low-level properties of the counterfeit and genuine tomatoes, or do we *misrepresent* the counterfeit tomato as a genuine tomato? This isn't an argument that we represent tomato-ness because we represent square-ness. Rather, my point is that substitution considerations won't settle the issue. Another auxiliary premise might appeal to information processing considerations. So let's turn to that argument as a way to secure the claim that perceptual content is restricted to low-level properties.

### 4.6.2 Information Processing Arguments

Information processing arguments—or their ancestors—have been around about as long as people have been theorizing about perception. For instance, in *On the Soul*, Aristotle provides us with a sketch of his theory of perception:

> The object of sight is the visible, and what is visible is color... Every colour has in it the power to set in movement what is actually transparent... it is only in light that the colour of a thing is seen... If what has colour is placed in immediate contact with the eye, it cannot be seen. Colour sets in movement what is transparent, e.g. the air... (Aristotle, 1984, 418a-419a)

Thus, we find in Aristotle the familiar idea that we only perceive surfaces properties. And the reason why he thinks "the visible" is restricted to surfaces has to do with

the way that light interacts with surfaces in the environment and our eyes. While the word "information" or its cognates do not figure into Aristotle's account, he nonetheless appeals to various properties of the stimulus to identify the proper "objects of perception."

Although I'm not interested in critiquing Aristotle's theory of perception, I include this passage in this discussion because it captures the predominant intuition that perception is somehow constrained by the properties of the light landing on our retinas. Indeed, as I pointed out above, the inverse optics problem concerns how our visual systems arrive at a (more or less) determinate three-dimensional model of our local environments from a two-dimensional light array landing on our retina. A natural way to conceive of that project is to assume that visual system's sole responsibility is to generate an inverse mapping—i.e., a mapping from sensory stimulation to the three dimensional world. Of course, the inverse optics problem shows that no such mapping can be derived from the sensory stimulation. But the prevailing thought is that with certain assumptions about how light tends to interact with the environment, the visual system reconstructs the distal objects that reflect the light. Since only surfaces reflect light, only those aspects of the environment are depicted.

This sort of view suggests that there's something about visual information that constrains what can be visually represented. One way of trying to flesh out the thought that facts about information processing constrain perceptual content is by appealing to "transduction." A number of authors flirt with this idea. Pylyshyn (2003), for example, says that the content of early vision is restricted to "transduca-

ble" aspects of the environment, those "properties whose detection does not require accessing memory and drawing inferences" (p. 163). In a discussion of iconic representation, Fodor (2008) says that perceptual icons can only represent "transducer detectible' properties" (p. 186).

Strictly speaking, the concept of transduction refers to processes that convert a signal from one energy form to another. For example, the cochlea converts compression waves into neural signals, by mechanically transferring the energy across a variety of structures in the inner ear, ultimately moving tiny hair cells that produce neural activation. Generally, the presumption is that transduction preserves the structure of the transduced signal (or at lease some of the signal's structure, some of the time). But whatever structure is preserved, it pretty clearly does not correspond to the structure of visual representations. This fact is entailed by the inverse optics problem. Thus, a strict conception of transduction doesn't, by itself, tell us that perceptual content is restricted to "transducable" because we have no reason to think transduction is a representational process.[8]

However, "transduction" can be used more liberally to mean a signal that (a) was transduced, and (b) has received some limited amount of processing. Indeed, Fodor (2008) uses "transduction" to refer to a process that "takes ambient energy onto mental representations" (p. 187). This understanding of transduction pretty clearly goes beyond the conversion of a signal from one form of energy to another. Thus, a plausible way of reading Fodor is that (strict) sensory transduction provides the input to a perceptual module, where the module contains a hypothesis space

[8]Fodor is quite aware of a more restricted notion of transduction. See Fodor (1983)

that is restricted to possible distal causes of the light hitting the retina:

> Sensory processes, according to this account, merely register such proximal stimulations as an organism's environment affords. It's left to cognitive processes—notably the perceptual ones—to interpret sensory states by assigning probable distal causes.[9] (Fodor, 1984, p. 36)

Notice on this proposal that it's the background theory that constrains which properties are represented.[10] We don't perceptually represent cloud chamber trails as protons because our visual system's background theories lack the term "proton." But on this way of thinking about the issue, it's an empirical question which terms comprise the visual module's background theory. Nothing about an information processing approach, itself, requires that the contents be restricted to low-level features of the environment. Indeed, Fodor (1983) suggests, on phenomenological grounds, that the outputs of perceptual modules corresponds to basic-level categories, such as CAR, DOG, HOUSE. So if we think that facts about the way the brain processes information will inform what the visual system represents, then we best look to evidence that will reveal what sorts of hypotheses are entertained by the visual system.

[9]This notion is very much at peace with Burge (2010) and Prinz (2006) discussed above, well as Fodor (1983; 1984)

[10]Some might find Fodor's talk of a module's "background theory" to be highly tendentious. As I intend the term, "background theory" is a placeholder term for whatever it is allows the visual system to perform an analysis of the incoming signal. This placeholder eventually needs to be replaced within an elaborated theory (an issue I'll be addressing in Chapter 6); but as long as we are careful not to over intellectualize the notion, we are free to posit some thing like a background theory within the visual system. Indeed, Bayesian theories of visual processing understand the system's background theory in terms of priors. Fodor's writing on this issue is really quite prescient of modern Bayesian theories of perceptual inference.

## 4.7 Conclusion

Summing things up, while there's substantial agreement that vision only represents low-level properties, this view is not well supported. On the one hand, substitution arguments are too strong, in that they seem to entail that we perceptually represent very little, if anything. On the other, information processing considerations are not, on their own, particularly informative. Whatever our visual systems transduce, that's not what we perceptually represent. Once we move beyond transduction, we enter the domain of inference (or, if you like "information processing"), and the question becomes what sorts of hypotheses does the visual system "entertain"? It's certainly plausible that the hypothesis space is constrained in various ways. But it's not clear why we should think that the visual system is restricted to hypotheses that concern only low-level features of the scene.

So what we're looking for, now, is evidence that would bear on high-low content debate. What sorts of facts constrain what we can perceptually represent? Fodor offers us a blueprint, I think, of how to go about doing this. We want data that bears on the nature of the hypotheses "entertained" by the visual system. Of course, we don't have direct access to these hypotheses, so we will need to infer the nature of the visual system from our best theories. I take up this issue in the following chapter.

# Chapter 5: Two Methods for Identifying Perceptual Content

The main conclusion from the previous chapter was that if we're serious about investigating the scope of visual content, then we have to think more carefully about what sort of evidence bears on this question. This chapter takes up and evaluates two methods for identifying perceptual content. I first address Siegel's (2010) "phenomenal contrast" method and argue that it fails to resolve disputes between low-level and high-level theorists. I then address Block's (2014) empirically-motivated proposal.

Block appeals to a set of experiments that demonstrate adaptation aftereffects in response to facial categories. Block takes this as evidence that we perceptually represent facial categories, such as emotion, race, and gender. These studies are interesting and important, and ultimately, I agree with Block's conclusion that we perceptually represent these categories. But Block's case rests on the claim that adaptation is an exclusively perceptual phenomenon. I argue that we ought to be highly skeptical of this claim, and therefore we should not accept Block's conclusions as they stand.

## 5.1 The Phenomenal Contrast Method

Siegel (2010) recently rekindled the high–low debate by arguing that we represent "kind" properties (*being a pine tree*, *being a tea cup*), the semantic properties of words, individuals (*being John Malkovich*), and causal relations ($x$ launched $y$). Her position, however, has not been without controversy. The main point of contention is her phenomenal contrast method, and what she thinks it shows.

The phenomenal contrast method attempts to manipulate one's visual phenomenology across different experiences while keeping the input the same. If there's a phenomenal contrast between two experiences, we can ask what best explains the difference. The assumption is that because the input is the same in both cases, one cannot explain the contrast in terms of a difference in low-level visual representations.

To get a feel for the method, consider the following example. Suppose we have an image to two men standing toe-to-toe staring each other down. All we see are their heads, but we might think they are fighters of some sort. Now suppose we altered that image so that it appeared as though one of the men is gingerly touching the face of the other. (Perhaps these two images are stacked on top of each other.) Most would agree that in the first image, the two men appear angry with one another. But the addition of the hand would seem to change the emotional attribute from anger to tender affection and/or lust.

Applying the phenomenal contrast method, we want to know two things: (a) whether there's a visual contrast between one's experience of the top and bottom

images (abstracting away from the low-level differences that the hand contributes), and if so, (b) whether the contrast is due to a perceptual representation of emotional attributes of the faces, and not merely the low-level visual representational differences.

At first blush the method has some appeal. It supposes, for instance, that a property represented in visual experience should make a difference to how things look. Furthermore, it doesn't directly rely upon introspection of perceptual content. That is, the method does not ask us to report the contents of our conscious visual experience. The only thing we are asked to introspect is whether there's a phenomenal difference between our experience of the top image and our experience of the bottom image, which is plausibly more reliable than attempting to directly introspect the contents of perception.

But there are serious problem with this approach, as well. For starters, we find disagreement about when we have genuine cases of phenomenal contrast. In our earlier example, do the faces appear different in the bottom image, or do we just think it does? The example strikes me as maximally conducive for generating the kind of phenomenal contrast that Siegel has in mind. And yet when I compare my phenomenology of the two images, I am entirely conflicted. On the one hand, yes, it seems there would be a difference. But on the other hand, the intuitions are ethereal. Surveying the literature on this issue, Fish (2013) notes that different authors commonly have diametrically opposed intuitions about these kinds of cases, and concludes that "it looks as though there can be serious clashes of intuitions as to whether there is indeed a phenomenal difference between any particular pair of

cases" (p. 47).

Second, even when we agree that we have a contrast between the two experiences, there are alternative explanations that don't require positing high-level visual content. One of Siegel's main examples involves acquiring a recognitional capacity to discriminate pine trees from other conifers. Siegel thinks it's plausible that one's visual phenomenology changes as one gains the ability to quickly and easily identify pines from non-pines. But do we need to posit high-level content to account for this change?

Consider an alternative explanation sketch that appeals to feature-based attentional effects. Knowledge about the diagnostic differences between coniferous trees might cause one to attend to different features of the tree. The pine tree novice might attend to the global shape of the tree, while the arborists automatically and unconsciously attends to the shape of the boughs, cones, needle clusters, or other diagnostic features. Because the observers are attending to the trees in different ways, they are each representing different clusters of low-level features. The fact that the two observers are visually representing different features could explain the contrasting phenomenologies, but without the need to posit high-level visual content.

Block (2014) offers an alternative way of explaining the initial intuition that there's a contrast between the two images. While there is a contrast in our overall phenomenology, the contrast is due to a difference in our *cognitive* phenomenology. Our beliefs about what is depicted produce a kind of "cognitive overlay": when viewing the top image, we believe that the two characters are violently at odds with

each other; when viewing the bottom image, we believe that the two figures are attracted to each other. If beliefs alter our overall experience, we have a potential way of explaining the contrast. On this proposal, one's visual experience is the same for each image. All that changes is the "cognitive overlay."[1]

Siegel (2010) addresses cognitive phenomenology as a possible alternative explanation of her contrast cases—what she calls "dwelling on a belief"—but rejects it on the basis of the following argument. Suppose Block is right that the reason for the phenomenal contrast between an arborist and a novice is that the arborist has a different non-perceptual belief (e.g., *That* is a pine), which gives rise to a different cognitive phenomenology. Now suppose the arborist views a faithful hologram of a pine tree and is then told that it's holograph. In this case, Siegel thinks that the arborists phenomenology of the holograph should be different from her phenomenology of a genuine pine tree, since we are assuming that a belief is responsible for the original phenomenal contrast. Yet, according to Siegel, it's plausible that the "hologram could look exactly the same as the genuine pine tree after you became an expert" (p. 105). That is, learning that one is looking at a *holographic* pine tree shouldn't alter how it looks.

I agree with Siegel's intuition about this case. If I am unknowingly looking at a holographic pine tree, and someone suddenly walks through the holograph, I don't expect that the holographic tree will subsequently look different (after, of course, the person passes through it). But the problem with Siegel's argument is that it

[1]The issue of whether we have cognitive phenomenology—i.e., whether non-perceptual states have a what it's likeness—is itself a disputed issue (Bayne and Montague, 2012; Carruthers, 2011). But I think Block's proposal is prima facie plausible.

changes the proposed explanation of the contrast midway through the example. Remember, the alternative explanation holds that the novice and the arborist share the same visual phenomenology; what is different about their overall experiences is their conscious beliefs or attitudes about the holographic pine tree. One can consistently agree with Siegel's intuition that the arborist's visual experiences of a real and holographic pine tree are identical—that both "visual objects" look the same—yet deny there's a distinctly visual contrast between novice and the expert cases. Indeed, an ardent conservative would hold that all three cases produce the same *visual* phenomenology when we hold constant the low-level features of the stimulus, positioning, gaze, and attention, even though the cognitive phenomenology might differ radically across these cases.

Now, Siegel has more to say about these kinds of objections to her arguments. My aim here isn't to conclusively demonstrate that her arguments are unsound, but to give a representative sampling of the difficulties facing her methodology.

Let us sum up these difficulties. First, it relies heavily on introspection about whether we find contrasting pairs of experiences. That we find disagreement isn't so much the problem. The problem is that we lack independent means of adjudicating the conflicts. Second, the method faces empirically motivated confounds in the form of attentional effects and cognitive phenomenology. Whether or not Siegel can address these concerns is an open question. But I think it's clear that there are serious challenges facing her approach to this problem.

## 5.2 Adaptation and Perception

So are there any better ways of addressing the perceptual content issue? Does vision science, for instance, provide a way of addressing this question that doesn't suffer from the same problems? Block (2014) thinks it does. He proposes that adaptation effects provide a way of identifying perceptual content. Very roughly, he think that "if it adapts, it must be perceptual."

### 5.2.1 Block's Proposal

Experiments that elicit an adaptation aftereffect have a fairly standard structure. Imagine a set of stimuli ranging along a particular dimension, such as light sources varying in brightness, rectangles varying in width, or lines varying in tilt. Participants are first asked to classify stimuli ranging along the given dimension, and then they are adapted (i.e., exposed, or repeatedly exposed, for some length of time) to a stimulus at one end of the stimulus spectrum. Participants are then retested on the first set of stimuli. In many cases, their latter judgements are systematically biased away from the adaptor. This is called a "repulsive" aftereffect. Take, for instance, the well-known waterfall illusion. Staring at a waterfall for 30 seconds will make stationary stimuli appear to be moving (paradoxically) upwards. The subsequent percept is biased away from the adapting stimulus.

A general property, then, of an adaptation aftereffect is that some stimulus parameter $p$ will produce a response $r$ prior to adaptation, and a different response $r'$ after adaptation. Often the aftereffect will be "repulsive," as in the waterfall

illusion. However, some adaptation aftereffects are "attractive," (or attractive under certain conditions) where the adaptation biases future responses towards the adapter stimulus. For example, if one adapts to a line oriented at 30° to the left, subsequently viewed vertical lines will appear tilted slightly to the right. However, if one adapts to a line rotated 60° to the left, a subsequently viewed vertical lines will appear tilted slight to the left—i.e., the aftereffect is attractive (Gibson and Radner, 1935; Clifford, 2002).

Adaptation effects are often explained in terms of a decrease of neural sensitivity, generally referred to as the "neural fatigue" model (Benucci et al., 2013). After prolonged stimulation, a neuron will respond less vigorously to its preferred stimulus during subsequent presentation. Neural fatigue gives rise to aftereffects because stimulus features are encoded by the aggregate activity of neurons sensitive to particular stimulus values. For example, suppose we have two neurons that are sensitive to direction of motion. $\text{Neuron}_\text{a}$ preferentially responds to downward motion and $\text{neuron}_\text{c}$ preferentially responds to upward motion. The direction-of-motion encoded by these neurons will be a weighted sum of their activity. In a preadaptated state, when the system is presented with a stationary stimulus, both neurons will fire at the same rate, therefore encoding that the stimulus is stationary. However, if we adapt this system to downward motion, $\text{neuron}_\text{a}$'s activity will be suppressed, effectively lowering the threshold for upwards motion. So when the system is presented with a stationary stimulus, it will encode the stimulus as moving upwards. This is the sort of model Block has in mind.[2]

[2]Researchers now think neural fatigue model is at best an over-simplification of some

A large literature now exists investigating adaptation effects in response to facial categories, such as gender, race, emotion, and identity. Block cites a study by Butler et al. (2008) who show that adapting to an angry face will make a neutral face appear frightened (and vice versa). Consider the faces in Figure 5.1. The left face appears angry, the right face appears fearful, and the middle face appears ambiguous between the two. If one "adapts" to (i.e., fixates on) the angry face for a prolonged period, the middle ambiguous image will appear fearful. (Readers can try this by covering up the right two faces and fixating on the left most face for around 30 seconds.) According to Block (2014), this experiment, and others like it, "grounds a prima facie case that we have visual attributives for facial expressions" (p. 5). The idea here is that one can infer the contents of perception from the sorts of things that induce aftereffects. That is, because paradigmatic perceptual contents (color, shape, spatial relations) induce aftereffects, we should expect other features that induce aftereffects to be perceptually represented.

*Figure 5.1: Images of emotional expressions. Left: angry; right: frightened; middle: ambiguous between angry and frightened. Adaptation to the left image will make the middle image appear frightened. Adapted from Butler et al. (2008) with permission from Elsevier.*

forms of adaptation, and an inaccurate model of others (Kohn, 2007; Clifford et al., 2007; Solomon and Kohn, 2014). My impression from the neuroscientific literature on adaptation is there is no single "adaptation" mechanism, but rather a variety of mechanisms that differ in important functional ways but nonetheless give rise to the same (or similar) psychophysical profile.

But as Block notes, the fact that facial features can induce aftereffects doesn't show that emotional facial attributes, as such, undergo adaptation (hence the "prima facie" qualifier). This is because we find adaptation to a variety of low-level stimulus properties, such as tilt, width, height, curvature of line and so on. So in the "emotional" adaptation experiments, participants might be adapting to a constellation of low-level features (e.g., the geometry of the eyes and mouth) that make up an angry face, which then produces an aggregate aftereffect. Block calls this the "recognitional coextension" issue.

Block's project, then, faces two theoretical challenges. The first is to tease apart which of the two contingently coextensive properties (or sets of properties) the visual system is representing (i.e., a set of low-level features that happen to comprise a face or an angry face, as such). The second is to explain why the facial adaptation effects are genuinely perceptual (i.e., that the aftereffects have a perceptual basis).

As for the former challenge, Block (2014) draws on studies that show that adaptation effects persist for facial expressions, even if the low-level properties are altered, "suggesting that face perception utilizes both low and high level attributives" (p. 5). This is an important point, as it shows how we can infer high-level or categorical content from the profile of psychophysical data. If low-level details are driving the adaptation effect, then altering the low-level details should extinguish the effect. But since the effect persists, the effect must be due to some abstract property of the scene.[3]

[3]I broadly agree with Block's line of reasoning. However, it seems to imply that abstract

Block cites a further reason to think that facial adaptation exhibits orientation-specific effects. Looking at the adaptation to elongated and compressed eye-nose-mouth regions, Susilo et al. (2010) find that "inverted face aftereffects are generated by shape-generic mechanisms, while upright face aftereffects derive from both shape-generic and face-specific mechanisms" (p. 13). That is, they find both low-level and high-level aftereffects, but only high-level face aftereffects for upright faces—which is what one would expect given that face processing is highly orientation sensitive (Aguirre et al., 1999). Thus, the orientation sensitivity of face adaptation provides indirect support for the hypothesis that emotional adaptation reflects high-level adaptation, as opposed to an aggregate adaptation.

Of course, it's debatable whether these results conclusively show that we have adaptation to high-level facial adaptation.[4] As Block (2014) notes, "no single result can rule out that the result is due to differences between low level features of right-side-up and upside-down faces (e.g., the downward rather than upward curve of the eyebrows)" (p. 7). The general line of thinking behind both studies is that the adaptation profiles for faces is organized along categorical dimensions, as opposed to merely low-level visual dimensions. I take Block's point to be that, even if this evidence isn't probative, it's nonetheless the sort of evidence that should bear on

properties *cause* the adaptation aftereffect. This raises a puzzle: how do abstract properties have effects on our perceptual systems? I want to suggest that they don't. I argue in the Chapter 7 that the visual system represents these properties not because the visual adapts to these properties, but because the visual system possesses a representational schema for facial emotions—a schema that organizes in the incoming sensory input along the dimension of emotions. On this view, abstract properties don't *cause* the adaptation aftereffect; rather a congery of low-level properties generates the aftereffect with an profile that ranges along emotional dimensions.

[4]Briscoe (2015) challenges the claim that facial adaptation is adaptation to high-level properties, though he doesn't consider the supporting evidence discussed by Block.

whether we perceptually represent high-level aspects of our immediate environment, as opposed to "armchair considerations."

While the recognitional coextension problem is interesting, and obviously important for Block's argument, the question that I want to explore is why we should think that the psychophysical profiles of these experiments have a perceptual basis. That is, assuming the psychophysical profiles exhibit a categorical structure, why should we think it's *perceptual* adaptation, as opposed to cognitive adaptation? It's not entirely obvious where Block stands on this issue.

Block seems to think that the adaptation effects reflect a change in how the image appears: "The face in the middle in Figure [5.1] looks first angry then fearful" (p. 4). Of course, if the only reason for thinking that this is a perceptual effect is because the images look different pre- and post-adaptation, then it's not clear in what sense Block's proposal is any less controversial than Siegel's. (Recall Fish's (2013) concern about conflicting intuitions of whether we find contrasting visual phenomenology.)

Block (2014) is sensitive to this worry: "But can we be sure from introspection that those 'looks' are really perceptual, as opposed to primarily the 'cognitive phenomenology' of a conceptual overlay on perception, that is, partly or wholly a matter of a conscious episode of perceptual judgment rather than pure perception?" (p. 7). To obtain the conclusion he's wanting, Block needs to provide independent considerations for why adaptation effects are perceptual in nature. Here, Block turns to a variety of findings in the vision sciences to bolster his case.[5]

[5]One note about terminology: Block talks of "concepts not adapting" in the way percepts do.

## 5.3 Block on Why Adaptation is Perceptual

Block begins by citing the role of adaptation in spontaneous perceptual reversals. Consider binocular rivalry. When two distinct images are projected onto each retina, one's conscious experience alternates between representations of the two stimuli. According to standard competition-based models, both images are processed in parallel, with each signal competing for dominancy. When a signal outcompetes the other, it maintains its dominancy by suppressing the signal originating from the opposite eye. It's thought that over time the visual neurons that encode the dominant stimulus adapt, decreasing their activity, as well as their ability to suppress the other stimulus (Alais et al., 2010). This allows the rival signal to increase in strength and eventually achieve dominancy. Thus, adaptation functions as a kind of "trigger" for binocular rivalry reversals. Since binocular rivalry is thought to be quite a low-level process, according to Block, this suggests that adaptation is perceptual.

A similar kind of story applies to spontaneous figure-ground reversals. When one perceives a face in the ambiguous face–vase image, the signals encoding the face shape suppress the signals encoding the (background) vase shape. After prolonged

It's not clear to me that this is the correct contrast. Recall that he holds that percepts are constituted by perceptual attributives and contextually specified particulars. Thus, we can characterize a percept of a moving waterfall (in an oversimplified manner) as "*That* is DOWNWARD MOTION," where "That" picks out a contextually specified (singular) visual object, and "DOWNWARD MOTION" is a perceptual attributive, assigning the property of downward motion to the visual object. When a perceiver adapts to downward motion, the "percept adapts"—that is, the "percept fixation" process is biased towards assigning "UPWARD MOTION" to visual objects. But if "percepts adapt," then the appropriate contrast in the cognitive domain is belief fixation or judgment. On this way of framing the issue, when asking whether "concepts adapt," we should be looking to see whether propositionalized judgments exhibit a characteristic adaptation aftereffect.

viewing, the face-shape signals fatigue, allowing the vase-shape signals to dominate (Leopold and Logothetis, 1999). Again, figure–ground segmentation is a paradigmatic perceptual process, so, as Block states, "at least some kinds of figure-ground alternations are perceptual, not conceptual" (p. 8).

Block also appeals to a study by Schwiedrzik et al. (2014) who show that the influences of adaptation and priming on subsequent percepts can be independently measured on a single task. These researchers found that prior exposure to a stimulus, $S_a$, of a particular orientation generated a repulsive aftereffect when participants were asked to identify the orientation of a subsequent stimulus $S_b$. However, they also found that *reports* of $S_a$'s orientation (whether the reports accurately tracked objective orientation) systematically biased subsequent reports of $S_b$'s orientation. If a participant reported $S_a$ being tilted to the left, then she would be more likely to report $S_b$ as tilted to the left. Note here that this bias has an attractive affect on perceptual judgments, which, according to Schwiedrzik et al., suggests that it's not adaptation, but rather priming.

Furthermore, using fMRI, they found that the priming effects are associated with frontal and parietal activity, while adaptation effects are restricted to purely visual areas (e.g., V2 and V3). This suggests, according to the authors, that priming affects participants' *judgments* of the orientation of the contours, while adaptation affects the *perceived* orientation of the contours. These studies nicely demonstrate that paradigm cases of adaptation are genuinely perceptual. Block will find no quarrel from me on this issue.

Block also points to the lack of existing empirical support for the idea that

adaptation is a cognitive phenomenon. He suggests that if adaptation were to occur in cognitive processing, then we should expect adaptation effects to manifest in the cognitive domain. One place where we might find such effects would be in cases of interpreting moral actions. We can imagine a set of relatively similar actions that range along a moral dimension, from morally abhorrent, to morally ambiguous, to morally virtuous. Adaptation to morally abhorrent actions would result in morally ambiguous actions being judged as morally virtuous. Block (2014) concludes by noting that "[N]o such phenomena have been reported to my knowledge" (p. 8).

The case of moral adaptation actually strikes me as an entirely plausible case of adaptation. Consider, for instance, the sorts of changes in moral attitude that occur in warfare, or the change in moral disposition exhibited in the Stanford Prison experiment (Haney and Zimbardo, 1998). Something like "moral adaptation" might figure into explanation of why American soldiers abused the prisoners at the Abu Ghriab prison. Perhaps the abuses began with rather harmless taunts and slowly progressed towards full-blown cruelty. At each step of the way, the soldiers' moral attitudes shifted to accommodate the new moral norm. This, of course, is entirely speculative. In order to conclude that cognitive systems adapt, we'd need much better evidence.

One point that Block oddly doesn't stress is the sheer ubiquity of adaptation throughout the sensory domain. We find auditory adaptation (Eatock, 2000; Pérez-González and Malmierca, 2014; Schweinberger et al., 2008), tactile adaptation (Musall et al., 2014), olfactory adaptation (Dalton and Wysocki, 1996; Sanchez-Vives et al., 2000), and vestibular adaptation (Young et al., 2003). In the visual domain alone,

standard cases of adaptation include (but not limited to) lightness (Palmer, 1999), color (Atlick et al., 1993), tilt (Clifford, 2002), curvature (Bell et al., 2009), and the spatial frequency of gratings (Kohn, 2007). We find adaptation at quite at peripheral sites of sensory processing (e.g., the retina) all the way to sites in neo-cortex (Webster, 2015).

However, the sheer ubiquity (and heterogeneity) of adaptation could also point to it being a *neural* phenomenon. The fact that we tend to find it in perceptual systems is because it happens to be easy to selectively adapt neural populations through perceptual channels. But if adaptation is a neural phenomenon, then we should see its affects in non-perceptual systems as well. In the following section, I argue that approximate numerical adaptation is a case of cognitive adaptation.

## 5.4 Approximate Numerosity: A Case of Cognitive Adaptation

So if we are looking for cognitive adaptation, what should we be looking for? I propose that we look for architectural markers of non-perceptual processing. If we find a system that (a) encodes abstract properties, (b) receives multi-modal input, (c) is independent of known perceptual systems, and (d) that system is responsible for a relatively well-defined adaptation aftereffect, then we should think that some cognitive content adapts. I argue that the approximate numerical system fits this description.

A variety of animals (humans included) have an evolutionary ancient approximate number system that allows them to make quick but "noisy" estimations of

numerosity. In humans, this capacity can be dissociated from higher-forms of discrete mathematical abilities (Dehaene, 1997). Psychologists can reliably probe this system by briefly presenting stimuli containing five or more items (e.g., an array of dots), such that the participants do not have time to count the items. Participants' responses (over many trials) follows Weber's Law, where the variance for small magnitude estimates is relatively small, and increases logarithmically as the magnitudes become larger.

Burr and Ross (2008) find that when participants visually adapt to a 400-item collection they subsequently judge smaller collections significantly larger than they would have normally. In other words, numerosity produces an adaptation aftereffect.[6] Interestingly, Burr and Ross (2008) conclude that approximate numerosity is a "primary visual property, like color or motion" (p. 425). Indeed, Fish (2013) cites this study to support the claim that we represent high-level properties in vision, and nicely captures the inference underlying this claim:

> Their reasoning behind this methodology was the observation that all agreed primary visual properties—the properties (such as size, orientation, shape, color and motion) that everyone agrees appear in phenomenal character—are susceptible to adaptation. So if we can show that another property is also susceptible to adaptation, we have an argument that this property appears in phenomenal character too. (p. 52)

So like Block, Burr and Ross (2008) and Fish (2013) assume that adaptation effects

[6] N.b., the experimenters controlled for stimulus density, orientation, and contrast (Burr and Ross, 2008).

are a mark of the perceptual.

However, on the face of it Burr and Ross's finding is as much evidence that adaptation goes beyond perceptual processing as it is evidence for numerosity being a primary visual property. It is generally accepted that approximate numerosity capacities are the result of a functionally distinct system. In support of this idea, one finds the same "approximate number line" (i.e., the same profile of approximate numerosity judgments) across a variety of stimuli and across multiple modalities (Feigenson et al., 2004). Indeed, people are no slower or less accurate when comparing numerosities across modalities than they are when making comparisons within a modality (Barth et al., 2003)

We also find corroborating fMRI evidence. Piazza et al. (2007) find that an area within the intraparietal sulcus responds preferentially to approximate numerosity (e.g., an array of dots). Importantly, however, this area also responds to other numerical symbols, such as Arabic, spelled-out, and spoken numerals. This suggests a common faculty for computing approximate numerosity, regardless of the format or modality.

Piazza and colleague's results are interesting not only because they suggest a common system for encoding approximate numerical quantity, but also because of the methodology employed in this experiment, namely an fMRI adaptation paradigm. The authors hypothesized that if a single brain area encodes approximate numerosity, then that area would exhibit adaptation effects across non-symbolic (arrays of dots) and symbolic (Arabic numerals) stimuli. For example, adapting to a 14-element array will reduce neural activity in the intraparietal sulcus overtime, and

a subsequent presentation of a symbol representing 46 will produce an activation "rebound" effect, which can be measured using fMRI.[7]

Note that this result merely shows a repetition suppression effect for numerosity. One might argue that mere suppression effects need not reflect adaptation, as understood as the characteristic repulsive aftereffect. Thus, it doesn't demonstrate the characteristic adaptation aftereffect found in psychophysical experiments. While I think we should be cautious about identifying all suppression effects with adaptation, repetition suppression effects are closely linked to aftereffects in that the selective suppression is thought to explain changes in perceptual processing (Grill-Spector et al., 2006; Krekelberg et al., 2006). Moreover, the neural suppression in the intraparietal sulcus region suggests that the suppression is driving the aftereffects. So the fMRI data provides considerable evidence that the numerosity system is responsible for the behavioral results.

Nevertheless, we can make a stronger case for numerical adaptation being a species of cognitive adaptation. Arrighi et al. (2014) provide evidence of adaptation aftereffects that generalize across modality and format. For example, adapting to sequences of lights produces an aftereffect when hearing sequences of tones, and adapting to sequential flashes produces an aftereffect when perceiving arrays of elements (simultaneous numerosity displays). Note that the cross-modal and cross-format adaptation eliminates many potential "low-level" confounds such as texture density, because there are few (if any) low-level features shared across modalities or

[7]N.b., the Piazza et al., (2007) results replicate earlier findings in Piazza et al. (2004).

format.[8] Arrighi et al. (2014) sum up their results as follows:

> That the adaptation [to numerical quantity] occurs across sensory modalities and across presentation formats shows that these separate ways of representing numeric information are highly interconnected, probably all feeding into one common representation of number. That cross-modal and cross-format adaptation effects were almost as large as within-modal and within-format adaptation suggests that it is the abstract quantity system that adapts, rather than the separate systems that feed it. (p. 5)

These results nicely complement the main conclusion of the fMRI adaptation studies mentioned above, showing approximate numerical judgments are the result of sensory-independent cortical structure. Further, the repetition suppression effects in the intraparietal sulcus offer a plausible explanation of the generalized numerosity adaptation.

But despite drawing these conclusions, Arrighi et al. (2014) think that numerical adaptation is a perceptual phenomenon. Their reason for thinking this is that they found that adaptation within the visual domain was spatially specific. That is, they found adaptation effects for numerical stimuli only when the adaptor and the test stimulus were in the same (objective) location in space (despite eye movements). When the adaptor and the test stimulus were in different locations, the adaptation

[8] Durgin (2008) argues that Burr and Ross's (2008) are better understood as adaptation to texture density, as opposed to numerosity. Although Burr and Ross controlled for some aspects of density, it's remains controversial if they adequately controlled for this confound. Arrighi and colleagues' results do not suffer from this potential defect.

effect disappeared. They take this to show that the adaptation effect is *perceptual* in nature, though "consistent with the adaptation occurring at moderately high levels of analysis" (Arrighi et al., 2014, p. 6)

As it turns out, this spatial selectivity result is not particularly compelling. What these researchers found was that if the adaptor and test stimuli are in different objective locations, but the same retinotopic location (i.e., the stimuli moved but participants' eye movements preserved the retinotopic location) the aftereffect was negligible. Adaptation was when participants' eyes moved, but the stimulus stayed in the same external location. In other words, numerical adaptation reveals a scene-based specificity, but not a retinotopic specificity.

What this result suggests is that numerical judgments track the numerosity of distal objects, as opposed to the numerosity of the proximal stimulus (the retinal projection). This, of course, is perfectly consistent with the hypothesis that numerosity is perceptually encoded. Visual spatial content is *anchored* to retinocentric coordinates (in the sense that movement of the eyes affects our perspective), but visual content is clearly encoded in scene-based coordinates as well. We perceive objects as having fixed locations, despite changes of perspective.

However, spatial selectivity to distal objects can be naturally accommodated on the hypothesis that approximate numerosity is a non-perceptual capacity. If approximate numerosity were a cognitive phenomenon, then we might expect adaptation to occur only in response to stable features of the environment. That is, we might expect adaptation to the numerosity of "things," as opposed to numerosity a raw sensation of numerosity. In this case, each numerical stimulus consists of

a sequence of light flashes, presumably represented as belonging to a single event or visual object. The visual system keeps track of the visual object, and sends numerically-relevant signals to the amodal approximate numerosity system. Indeed, this is precisely how Dehaene characterizes his theory:

> The occipito-parietal areas [visual areas] of the brain contain neuronal ensembles that rapidly extract, in parallel across the visual field, the locations of surrounding objects. Neurons in these areas seem to encode the location of objects regardless of their identity and even to maintain a representation of objects that have been hidden behind a screen. Hence, the information they extract is ideally abstract to feed an approximate accumulator. (p. 68-69)

The "accumulator" is the hypothesized mechanism that keeps track of analogue numerical content. Piazza et al.'s hypothesis is correct, the intraparietal sulcus is the neural basis of this accumulator. But if the accumulator is distinct from the visual system, then the numerical content is represented independently of the visual system. These considerations, of course, don't demonstrate that approximate numerosity must be cognitive; rather, they demonstrate that spatial selectivity does not, on it's own, tell us whether the adaptation is perceptual.

On balance, then, the evidence suggests numerical adaptation occurs in a system that is distinct from the visual processing. We have independent evidence of a distinct numerical system that activates in response to spoken and written numerical symbols, and to auditory and visual (non-symbolic) numerical stimuli. And this

suggests that the most parsimonious explanation for Arrighi et al.'s adaptation results is that a vision-independent system undergoes adaptation.

At this point, it would be worthwhile to take stock of the argument. I have been arguing that adaptation goes beyond the perceptual domain because we find adaptation to approximate numerosity. We should think that approximate numerosity system is cognitive, as opposed to perceptual, because (a) it represents numerosity independent of modality (or at least sight and audition), (b) it represents numerosity independent of format, and (c) numerosity is an abstract property.

On balance, then, the evidence suggests numerical adaptation occurs in a system that is distinct from the visual system, given that we have independent evidence for such a system. Thus, we ought to resist Block's urge to treat adaptation aftereffects as a reliable guide to perceptual domain.

## 5.5 Conclusion

I have argued in this chapter that both Siegel's phenomenal contrast method and Block's adaptation-based method ought to be viewed with a healthy dose of skepticism. Siegel's approach relies quite heavily on introspection, and for this reason I don't see it gaining traction on issue of visual content. Block's appeal to findings in the vision sciences is a step in the right direction. If we are to make progress on this question, then we need to avail ourselves of the rich theoretical frameworks in the vision sciences.

Although I am critical of Block's argument for why we perceive high-level

facial emotions, I nonetheless agree with this conclusion. In the following chapter, I layout a framework in which to address the question of perceptual content. This sets up Chapter 7, where I argue that we represent high-level properties.

# Chapter 6: Organizing Principles and High-Level Content

In the previous chapter I argued that Block overstates the case for thinking that adaptation is exclusively perceptual. And if the locus of the emotional adaptation effect is cognitive, then we cannot conclude that emotions are perceptually represented. I propose that if facial adaptation is a perceptual phenomenon, then there needs to be some sort of abstract representational structure embedded within the visual system that explains the aftereffects. How might we go about identifying such representational structures, and how can we determine whether they are perceptual in nature? To address this question, I suggest we look to how vision scientists explain our visual capacities.

In this chapter, I argue that organizing principles are a fundamental posit in vision science. Very generally, we can understand their role as imposing a particular organization upon the incoming sensory stream. It is in virtue of these organizing principles that we perceive basic aspects of the environment, from contours to bounded three-dimensional objects. So, I propose, if we find evidence of visual principles that organize the input along abstract dimensions (as opposed to basic geometric and/or chromatic dimensions), then we can conclude that we represent high-level, abstract properties.

## 6.1 Two Kinds of Visual Representations

In her discussion of the explanatory frameworks in the vision sciences, Orlandi (2014) notes that theorists tend to distinguish between two kinds of visual representations. Following her terminology, we have sensory representations (s-representations) and representations of principles (p-representations).

We can think of s-representations as the standard set of sensory content fixations performed by the visual system, such as luminance changes, orientation, motion, and color. When we see a red square, our visual system contains an s-representation of the redness and an s-representation of the squareness. Our overall visual experience of the red square, then, consists of a unified s-representation.[1]

P-representations, then, are "the rules encoded in visual systems" that "regulate the way percepts are derived from sensory states (Orlandi, 2014, p. 21). The basic idea is that the visual system embodies various organizing principles that specify how percepts are constructed from the incoming sensory signals. Unlike s-representations, p-representations do not figure explicitly into the unified s-representation (or percept). A unified s-representation is typically about the various things in our environment—the tables, chairs, curtains, etc. If I s-represent the greeness of a jade plant, I typically experience this greeness. P-representations work behind the scenes, making it possible to see the greeness of the jade plant.

Theorists often appeal to the ambiguity of the retinal stimulus to justify posit-

[1]Note that while s-representations are often conscious, they needn't always be. My visual system may s-represent a red patch on the wall without me being conscious of it. I bring conscious perception into this discussion merely to motivate the distinction between s- and p-representations.

ing p-representations. For example, a circular retinal image with radius $r$ might be a small circular object viewed close up or a large circular object viewed at a distance. Because we can't "directly detect" circle size—i.e., there is no way to encode the size of a circular object from its projection onto the retina—we must infer its size on the basis of other visual cues and background visual "knowledge" about the relationships between these cues and object size.

While I think that the proximal stimulus is ambiguous and requires that we posit internal procedures and knowledge, note that whether or not the stimulus is ambiguous is really beside the point. For even if the proximal stimulus were not ambiguous with regard to the size of the object—i.e., even if there was a one-to-one mapping from the proximal stimulus to the size of the object—we would still need to posit some sort of psychological structure or process to utilize this mapping. Consider the following analogy. The addition function maps two-tuples of the natural numbers onto a one-tuple of the natural numbers. Now suppose an agent contemplates a set containing two items and a set containing three. This agent can't conclude that there are five items in total without implementing the addition function (excluding, of course, lucky guesses). So even if there's a precise mapping from one domain to another, in order to make use of that mapping, an organism requires the appropriate psychological structure.

This last point speaks to a broader Kantian (or anti-Lockean) theme, which is that we require some kind of internal psychological structure in order to perceive or comprehend the world. Our minds are not blank canvases on which sensory impressions are imprinted. Undoubtedly, much of this structure is innate, and sub-

sequently shaped by brute causal interaction with environmental factors. Other structures may be learned. The origins of these structures doesn't much matter for our purposes. What matters is the recognition of the need to posit internal structure to explain perceptual capacities.

## 6.2 P-representations as Organizing Principles

To get a sense of the role that p-representations play in theories of perceptual processing, consider the depth-from-shading example in Figure 6.1. The top right and bottom left disks appear as convex bumps, while the top left and bottom right disks appear as concave dimples. The only difference between the bumps and dimples is that the bumps have shading on the bottom portion of the circle and the dimples have shading on the top portion of the circle. This simple shading cue is enough to generate a visual sense of depth.

The prevailing theory about why this occurs is that the visual system encodes the fact that objects tend to be lit from above, a fairly reliable assumption given that the sun is the primary source of illumination. Of course, objects aren't always lit from above. The illusion of concavity and convexity in Figure 6.1 likely reflects a strong bias to assume that objects are lit from above. But this bias can be overridden in the presence of countervailing visual evidence (Thomas et al., 2010). If one adds further structural and/or shading cues that indicate that the light source is from below (say, by adding a protruding cylinder that casts a shadow upward), one's

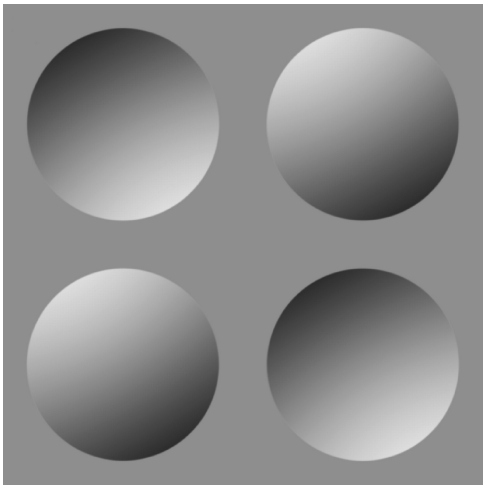

*Figure 6.1: An example of depth from shading. Because the visual system, by default, assumes objects are lit from above, shading on the top portion of a disk is consistent with it being a shadow cast from the edge of a concave dimple, while shading at the bottom portion of the disk is consistent with it being a shadow cast by a convex bump. Hence, the left-top disk appears concave while the top-right appears convex. From Stone (2011), with permission.*

perception of depth will be reversed. Essentially, if you know the rough location of the light source, the shading/shadow tells you a fair bit about the shape of the object. Positing that the visual system encodes this relationship allows theorists to explain, at least in a preliminary way, why some of the circles appear convex and others appear concave.

The idea that the visual system encodes the relationships between lighting and form is the idea that the visual system possesses, and utilizes, organizing principles (or p-representations) when constructing a visual percept. There should be nothing particularly controversial about the idea of visual organizing principles thus far. They are standard posits within the vision sciences. According to David Marr (1982), attempting to understand the psychological structure of the visual system is a central goal of vision scientists: "the business of isolating constraints that are both powerful enough to allow a process to be defined and generally true of the

world is a central theme of our inquiry" (p. 23).

More controversial is how best to understand the fundamental nature of organizing principles embodied in the visual system. For example, it's unclear whether visual organizing principles are always, never, or sometimes representational. Although Orlandi introduces the terminology of "s-representations" and "p-representations," she in fact denies that the visual system is representational (with one caveat that I will mention presently). That is, she rejects classical computational/representational theories of mind, which hold that visual processing consists of (in Orlandi's terms) explicitly encoded rules performing operations over sensory representations. She argues that p-representations are better understood as constraints wired into the architecture of the visual system, and s-representations are better understood as "tracking states"—states that carry information about a distal stimulus, but lack veridicality conditions. The one caveat is that she doesn't deny that the output of visual processing is representational; she merely denies that the mechanisms responsible for producing this output are representational.

While I am sympathetic with Orlandi's concerns about over-intellectualizing visual processes, and I tend to agree that some visual processes are non-represntational in the sense she describes, I think most organizing principles are representational in that they are decoupled from the proximal sensory causes. Recall our depth-from-shading example. This principle encodes the relationship between shading and direction of light to produce an estimation of depth. The visual system is in causal contact with the sensory input. It responds to various cues in the proximal stimulus, which are caused by reflectance properties of the distal objects. But it outputs an

estimation of depth, which is "contained" neither in the proximal nor distal stimulus. That is, the outputs of the organizing principles are not *about* the typical causes that generate those outputs. It's in this sense that the organizing principles are decoupled from the environment.

But even if one is a representationalist about organizing principles, there remains a variety of theoretical options to choose from. One might think the principles are explicitly encoded or one might think they are implicitly encoded. The explicit–implicit distinction is difficult to characterize in non-circular terms. But we can think of both sorts of encodings schemes as giving rise to rule-governed behavior. Very roughly, the difference lies in the fact that an explicitly encoded rule is one that is encoded as a string of symbols or some other kind of data structure, while an implicitly encoded rule is part of the structure of the hardware or wiring.[2]

There are also competing mathematical frameworks for describing visual processing. Classical computational approaches posit rules which constrain the possible interpretations of the sensory information (Marr, 1982; Pylyshyn, 1999). Bayesian theories of perceptual processing posit "priors"—probabilistically encoded assumptions or hypotheses—that are tested against the sensory evidence (Clark, 2013; Hohwy, 2014). One important difference between these approaches is that different rule-based constraints involve different computational operations, while Bayesian operations are fundamentally the same. For example, a rule-based constraint might say "if shading is below a contour, treat area above contour as convex, unless x,

[2]Orlandi views (or seems to view) any implicit encoding as non-representational. I remain neutral on this issue.

y, z..." Here, the computations would be performed over formally defined symbols that correspond to "shading," "contour," etc. However, Bayesian computational operations always involve updating prior assumptions based on the incoming sensory signal or "evidence." The fundamental computational procedure is always the same.[3]

For the most part, I wish to remain neutral on the computational properties of organizing principles. This is for a couple of reasons. First, all that my view requires is some sort of psychological structure that constrains the interpretation of the incoming sensory signal, and both classical computational and Bayesian approaches endorse this basic idea. My second reason for remaining neutral between these two computational frameworks is that it's not clear to what extent these two theoretical frameworks are at odds with one another.

On the one hand, classical rule-based theories require some sort of constraint satisfaction procedure for their application that allows for the flexible application of the rules. Suppose we have a default "light-from-above" rule used to interpret shading cues. Because light does not always fall from above, this rule must be flexibly applied, otherwise we would regularly misperceive convexity for concavity (and vice versa) whenever the visual scene is lit from below. Thus, the appropriate application of a direction-of-light rule needs to satisfy a number of constraints (e.g., apply light-from-above rule only when certain visual cues are not present—cues indicating light from below). Spelling out how this works could involve assigning weights to different kinds of visual cues. When there are sufficient cues that objects are lit from below,

[3]Thanks to Eric Mandelbaum for bringing this issue to my attention.

this evidence would "defeat" or override the "light-from-above" rule, and invoke the "light-from-below" rule. Such a constraint satisfaction procedure could make liberal use of Bayesian inference.

On the other hand, Bayesian theories are largely silent about the nature of the encoded hypotheses. Bayesian hypotheses are typically understood as propositions—"the world is such and so." But we could plausible construe talk of "hypotheses" in terms of rules or procedures about how to integrate sensory information. Instead of encoding "objects tend to be illuminated from above" the "hypothesis" might consist of a procedure for utilizing shading information in a particular way. The Bayesian formalism would be the same, but the psychological structure of the hypothesis would differ. But this sort of view appears identical to the rule-based view when one allows for the flexible application of the rules. Thus, at the outset, we can plausibly envision a happy marriage of rule-based and Bayesian theories of visual processing. Whether such a marriage is in the cards is not an issue I will address. My only point is that both proposals require elaboration; and since such a project is beyond the scope of this work, I shall henceforth remain neutral on the issue.[4]

I have made two points in this section. First, I have identified the role that p-representations play in standard theories of visual processing, viz., to constrain pos-

[4]Another reason for construing organizing principles broadly concerns the wide variety of phenomena p-representations are invoked to explain. Simple Gestalt grouping principles (e.g., proximity: elements close together are grouped together), modal completion (perceiving edges where there are none), color constancy (seeing a surface as the same color despite radical differences in the surface reflectance properties, feature/sensory integration (the integration of various streams of sensory information) all presuppose some sort of pre-existing psychological structure or rules. Indeed, in later sections of this chapter and the following one, I argue that object, causal, and animacy perception involve some sort of representational structure, but it's entirely plausible to understand simple Gestalt rules as resulting from the way neural populations are "wired" together. We might be able to subsume all these phenomena under a single theoretical framework. But I see no reason to assume this can be done at the outset.

sible interpretations of the sensory information. Second, I have noted a variety of different theoretical commitments one might hold with regard to "p-representations." I argued that organizing principles are decoupled from their distal causes, and are therefore representations. I choose to remain neutral with regard to the computational status of organizing principles, as both are minimally committed to positing psychological structure to explain perceptual capacities, such as the ability to perceive depth from the shading cues in 6.1.

## 6.3 Organizing Principles and Visual Content

We can think of organizing principles in two ways. On the one hand, we can think of them as static entities (e.g., collections of rules or abstract psychological structures). As static entities, organizing principles can be understood as p-representations—i.e., as having content. Second, we can think of organizing principles as processes. As processes, organizing principles play an active role in interpreting and organizing the incoming sensory signals. I argue that, in organizing the incoming sensory signals, organizing principles play a constitutive role in fixing percepts (or s-representations). To get the sense of my claim, here, consider the following two examples: Gestalt grouping principles and lightness constancy.

I assume that nearly all viewers see the dashed line in 6.2 as consisting of two intersecting curves. Of course, there are other interpretations that are consistent with the stimulus. A creature whose visual system embodies quite different organizing principles might see two curves meeting at their apexes, or four curves meeting

at their ends, and so on. Standard explanations for why we see the stimulus in the particular way we do appeal to the Gestalt principles (Palmer, 1999). In this case, the principle of proximity groups elements that are relatively close together, forming contours.

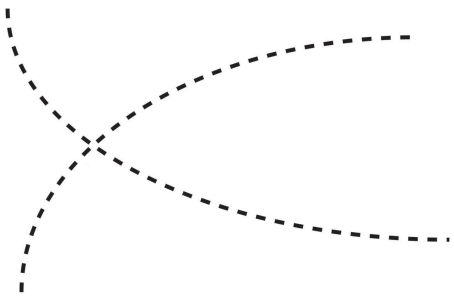

*Figure 6.2: An example of Gestalt grouping principles of proximity and good continuation. (Original drawing)*

But grouping by proximity doesn't solve the primary ambiguity problem in this picture, as the visual system needs to resolve which contours segments belong to the same curve. Here is where the principle of "good continuity" comes in. The principle of good continuity groups elements (or sets of elements) that share continuity without abrupt changes in trajectory. Because the top left and bottom right curve segments share continuity in this sense, they are grouped together, and the same for the top right and bottom left segments.[5] So for our simplified example, the organizing principles of proximity and good continuity produce a representation of two crisscrossing lines.

Gestalt grouping principles used to generate perceptual representations of contours are but one small example of where we find organizing principles in the vision

[5]It should be noted that, while this is at best a sketch of an explanation, considerable work has been done to formalize and operationalize Gestalt grouping principles (Wagemans et al., 2012). For our purposes, a rough-and-ready description of Gestalt grouping principles will suffice.

sciences. Visual context effects offer a number of good examples. Color and brightness constancies arise when the perceived color or brightness of a stimulus is invariant to the color or brightness of the stimulus stimulus. Consider the case of brightness contrast in Figure 6.3. The squares labelled A and B appear to be distinctly different shades of gray even though they have identical reflectance properties. The visual system utilizes various contextual cues to determine the relative shades of each square, such as the local contrast differences of the individual squares and the fact that B is in the cylinder's shadow. While it remains an open theoretical question how the visual system "knows" that B "has less light falling on it," one low-level cue is the nature of the shadow's border.[6] In normal environments, light emitted from a single source scatters, such that the edges of a shadow are fuzzy. Fuzzy contrast changes (as seen in the squares surrounding B) suggest that the contrast is the result of a change in illumination, as opposed to surface properties (Purves and Lotto, 2011). Of importance for the present issue is that various contextual contrast properties must be somehow integrated to generate a local determination of lightness. Clearly, something needs to orchestrate how various contextual cues affect the perceived lightness of a particular square. This phenomenon calls out for some sort of organizing principle or set of principles.

My view is that organizing principles play a constitutive role in fixing visual content, *because* the organizing principles play an essential role in explaining the visual percept. That is, the particular nature of the principles explain why we vi-

[6]Of course, because this is an image, the cylinder doesn't actually cast a shadow. If useful, the reader can imagine she is looking at three-dimensional set of objects.

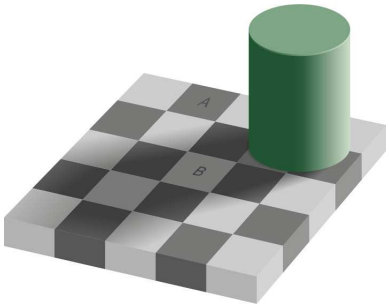


*Figure 6.3: An example of contrast constancy. The squares labeled A and B appear different shades of grey even though they have identical reflectance properties. (Image by Edward H. Adelson. Reproduced here with permission.*

sually parse the scene one way and not another. If our visual system possessed different grouping principles, we would perceive a different organization—i.e., the content of visual perception would be different. Going back to the Gestalt example, the principles of proximity and good continuity explain why we see the crisscrossing lines, as opposed to other possible configurations. If we had visual systems that embodied different psychological structures, then we would see the figure a comprising a different organization, and we would visually represent the scene differently.

While any particular organizing principle should explain why we see $X$ and not $Y$, organizing principles also explain why we represent the figure as having any sort of organization whatsoever. In my Gestalt example, it appears our visual system embodies something like the principle of common fate, in conjunction with luminance contrast "detectors." But the more general point is that we require some sort of mechanism that embodies a principle or set of principles that allows us to organize the incoming sensory information. Without a set of organizing principles we would be unable to make visual sense of the scene.

Consider a linguistic analogy. Without the ability to parse the syllables and words of a particular language (or dialect), utterances of that language are perceived

as a *linguistically* undifferentiated stream of noise. When we learn a language, particularly at a young age, we develop capacities that allow us to organize the incoming sound stream into linguistic components. Likewise, without the ability to parse the incoming visual signal into features and things, we would be unable to make visual sense of the energy stimulating our sensory receptors.

The general point is that organizing principles play a "Kantian" role, in that they make it possible for us to see the world as anything more than a fleeting and incoherent stream of sensations. Because organizing principles make it possible for us to see aspects of the scene as lines or in depth, they confer content onto the outputs of their operations. Thus, understanding the kinds of constraints that organizing principles impose on the incoming sensory stream provides important evidence for determining the content of visual processing. The examples that I discussed so far concern organizing principles for low-level features of the scene. In the following section I argue that object perception requires quite abstract organizing principles.

## 6.4 Object Perception

In this section, I will highlight the role of organizing principles in theories of object perception. The empirical literature on object perception is vast. My aim is not to provide anything like an exhaustive survey of the field, but rather motivate the need for abstract organizing principles in a domain that is generally considered perceptual. I begin by explicating the notion of object files and the motivations for positing them. I then briefly discuss some of the broader explanatory aims of

theories of object perception.

A first approximation of the goal of object perception is to explain how the visual system segments the scene into discrete entities—that is, how visual mechanisms utilize information in the light array to identify edges, surfaces, colors, textures, and then use this information to construct a representation of bounded entities.

One influential framework for theorizing about this problem posits "object files." Object files are "temporary 'episodic' representations of real world objects" and are distinct from "representations stored in a long-term recognitional network" (Kahneman et al., 1992, p. 176). The notion of an object file grew out of Treisman's "feature-integration theory" of attention. According to this model, different visual features (e.g., color, shape, orientation, spatial frequency, direction of movement) are processed in parallel by distinct subsystems. But if visual features are processed independently of one another, how is it that we perceive objects as unified entities?

This theoretical issue is called the "binding problem" and according to the feature-integration view, "focal attention provides the 'glue' which integrates the initially separable features into unitary objects" (Treisman and Gelade, 1980, p. 98). Although the view that attention is either necessary or sufficient for feature binding (that attention is the "glue" that binds features) is widely rejected, vision scientists continue to embrace object files for a variety of reasons. One reason is that object files offer a plausible framework in which to address the binding problem. Something needs to bring together information processed in disparate areas of the brain in order to produce a unified percept of the various objects in our environments. Object files provide a plausible framework in which to conceptualize

this problem.

A related motivation comes from research on the "object-specific preview effect." This paradigm involves participants viewing a display (the preview display) that contains two or more letters inside "visual objects" (e.g., boxes or triangles), and then identifying a single letter in the subsequent display (the target display). In some versions of the experiment, the visual objects change position, giving the impression of movement. The main finding is that participants are quicker to identify letters when the letters occupy the same visual object as they did in the previous display (as opposed to a different visual object)(Kahneman et al., 1992).

But why is it easier to identify a letter when it is contained within a two-dimensional shape? Here, object files offer a tentative explanation. This response time benefit arises because, upon viewing the initial preview display, the visual system "opens" object files for each of the objects. (The file will be subsequently closed shortly after the object goes out of sight.) If an object contains a letter, that information will be stored in the object file. So even if the letter moves location, the object file remains open and continues to store that information. When called upon to identify the letter in the target display, cognitive systems have ready access to the information in the object file, thus facilitating recognition of the letter.

Given that object files are thought to bind various sensory contents into a unified whole, it's tempting to assume that representing an object (more or less) involves assigning various sensory features to particular locations in perceived space. For example, according to a "simple binding" view, if you are looking at a red apple, your visual system assigns ROUND and RED to a particular region of space.

Your visual representation, then, will simply consist of those contents bound to a particular location: ([ROUND, *there*] & [RED, *there*]). But this implies (or seems to imply) that the identity of the representation is determined by the sensory contents bound into it. Hence, the representation ([ROUND, *there*] & [RED, *there*]) cannot be the same as ([ROUND, *there* & GREEN *there*]).

However, theorists dating back to Descartes (1984) have noted that we see objects as persisting through time despite manifest alterations to appearances. Descartes observed that one can alter the shape, smell, texture, density of a piece of wax, but we nonetheless identify the "substance" as the same. Thus, something besides superficial perceptual properties must account for idea of enduring bodies or substance—namely, a prior, innate idea of substance.

But this poses a puzzle for the simple binding view. If object representations are individuated in terms of the content bound into them, we would predict that any time an object changed color, we would see it as a different object. Thus, the simple binding view fails miserably to account for a central datum of object persistence—namely, the fact that objects persist despite changes to their appearances.

This observation inspired Pylyshyn to develop a theory of visual demonstratived. Building upon Kahneman et al.'s (1992) theory of object files, Pylyshyn (2000; 2001; 2007) argues that the visual system embodies a "primitive" indexing mechanism that tracks visual or "proto-objects" as they move through visual space. He calls this mechanism "Fingers of Instantiation" or FINSTs. According to Pylyshyn, FINSTs directly track (i.e., without representing) the location of up to four objects as they move through space, and these trigger the opening of an object file for each

object tracked. His motivation for positing such a mechanism is as follows:

> The theory is based on the recognition that to allocate focal attention, or to do many other visual operations over objects in the visual scene (e.g., encode their spatial relations, move focal attention to them), it is first necessary to have a way to *bind* parts of representations to these objects. A less technical way to put this is to say that if the visual system is to do something concerning some visual object, it must in some sense know *which* object it is doing it *to*" (p. 201).

The basic idea is that there needs to be some means of picking out objects that is independent of the sensory contents that are bundled together to form a unified representation of an object. The reason for thinking this is rooted in the Cartesian observation noted above, namely, the sensory appearances of objects do not remain constant through time, yet we perceive objects as persisting. A representation of an object through time cannot simply be a sensory description of the object. Pylyshyn argues that visual indices individuate object file representation. If he's right, then object individuation does not require an organizing principle.

Pylyshyn describes a number of "multiple object tracking" studies to support his visual indexing theory. A typical object tracking display consists of number of individual "objects" (two-dimensional shapes). While at rest, some subset of these shapes are singled out by (e.g., by a flashing the two-dimensional shapes) and participants are asked to track those objects (the target objects) during the subsequent stage of the experiment, where all objects move unpredictably about

the display. During the final phase of the experiment, the objects cease moving, a single object is picked out by an arrow, and the participants are asked whether it was one a member of the target group. Participants reliably give the correct answer for target sets containing up to five objects.

Importantly, Pylyshyn (2003) argues that "surface features" are not used for tracking objects. He notes that in several experiments, all of the objects in the display share the same features (e.g., all are black circles), and therefore "observers could not have been keeping track of the targets by using a unique stored description of each object, since at each instant in time the *only* property that is unique to each object is its location" (Pylyshyn, 2003, p. 226). That is, a feature-based tracking mechanism could work only if the various objects possessed different features. Since the objects do not differ in this way, Pylyshyn takes this as evidence that the visual system picks out these proto-objects demonstratively by tracking (but not representing) an object's spatiotemporal path.

There are a number of subtle issues bound up in Pylyshyn's discussion, namely, whether a visual index theory can be fleshed out without appeal to representation, and whether sensory features are used to track objects. For our purposes, I want to set aside the issue of representation. It may be the case that the visual system possesses a primitive mechanism that tracks spatiotemporal paths of objects, and this system is non-representational. Indeed, many studies suggest spatiotemporal cues play a predominant role in explaining how we track objects (Scholl, 2007).

However, Pylyshyn errs in claiming that only spatiotemporal contiguity is used to individuate objects. Papenmeier et al. (2014), for instance, find that surface color

affects multiple object tracking when spatiotemporal information is unreliable. In one experiment, an animation depicts spherical objects as though there were moving across a floating surface. (Imagine marbles rolling around a game board at table height just a few feet in front of one's body.) So as opposed to objects being depicted as moving across a plane (as in the case of standard object tracking displays) objects are depicted as moving through all three dimensions. As the objects move around the surface, the surface rotates, thus making it difficult to use spatiotemporal contiguity to track the objects. Indeed, when objects are of uniform color and size, rotating the scene impairs tracking performance. However, Papenmeier and colleagues found that performance is better when the target objects remain the same color throughout the trial, as opposed to changing color.

A priori, this result makes sense. Tracking objects in the real world involves tracking them as they move through all three dimensions. But because the retinal image is two-dimensional, we get an ambiguity problem: different sets of object trajectories can give rise to the same retinal stimulation. In addition to the inherent ambiguity of the retinal image, changes in viewer positioning can dramatically increase the difficulty of using spatiotemporal information alone to track objects. Different three-dimensional trajectories might map onto distinct two-dimensional paths, but they may be quite similar, and therefore difficult to discriminate between. And so one would expect that any reliable information about object identity would be used by the visual system. Because object color tends to remain stable over time, it provides an important clue about object identity.

Of course, the application of this information needs to be flexibly utilized, oth-

erwise we couldn't track objects that change color. Hence, Papenmeier et al. (2014) offer a "flexible-weighting" model, according to which "spatiotemporal information and surface features are both used to establish object correspondence and weighted according to the reliability of available information" (p. 168). The basic idea is that the visual system uses a variety of different cues to track objects and the extent to which a cue bears on content fixation is determined by prior visual knowledge and visual context. The point here is that, although spatiotemporal information is clearly used for tracking, it's not the only feature that is used.

Furthermore, we find informal evidence that spatiotemporal contiguity is not necessary for object individuation. Pylyshyn and Scholl's work on object individuation primarily observes conditions under which individuals track multiple moving objects. Hence, spatiotemporal contiguity is a predominant factor in explaining individuals' performance on this sort of task. However, as Brovold and Grush (2012) note, we find object individuation without spatiotemporal contiguity. Consider their example of a visual object defined by Gestalt grouping principles in Figure 6.4. The object here is stationary and spatially discontinuous, yet we see it as a pineapple. Brovold and Grush (2012) argue that because we find perceptual objects defined without spatiotemporal contiguity, there must be something more fundamental than spatiotemporal contiguity. They suggest that all object individuation involves Gestalt grouping principles, where spatiotemporal contiguity is an instance of Gestalt grouping principles that group features across time.

Thus, Pylyshyn's view that objects are individuated solely by spatiotemporal

*Figure 6.4: A "partially occluded" pineapple demonstrating that features other than spatiotemporal contiguity can be used to define visual objects. (Original drawing.)*

contiguity cues is at odds with the empirical evidence. However, there's something right about his insight which motivates his visual indexing theory. Pylyshyn is primarily resisting views that attempt to explain object individuation in terms of a mental description. That is, Pylyshyn is resisting "descriptivist" views of reference, of the sort advanced by Russell (1910) and Frege (1948): we represent object A as distinct from object B because the object file for object A contains a different description from object B. (We can think of the descriptions as lists of sensory feature, such as color, texture, shape, etc.) But, as Pylyshyn points out, because the object files of A and B can contain identical "lists" of features, one cannot appeal to different object file contents to explain how the brain distinguishes the two objects. This leads Pylyshyn to posit a purely causal tracking of spatiotemporal contiguity, a kind of direct reference theory of object individuation. But since this view can't be right, we are left with an apparent dilemma. Let us call this Pylyshyn's dilemma.

One way out of this dilemma is to distinguish between the contents of ob-

ject files, on the one hand, and their creation and maintenance, on the other (Gao and Scholl, 2010).[7] Doing so allows us to ask two independent questions: first, what sorts of sensory contents are bound into an object file; and second, what sorts of visual information can trigger the creation of an object file and govern its maintenance. By framing the issue in this way, we stipulate that representing two objects as distinct entities involves the construction of two distinct object files. That is, two objects are represented if and only if two object files are constructed.[8] This allows for two object files to contain identical sensory content, yet be distinct mental entities. Once we allow for distinct object files for distinct objects, the objects are represented (by definition) as distinct entities.

The real explanatory work for a theory of object individuation, then, concerns how object files are constructed and maintained. This requires a decision procedure along the lines proposed by Papenmeier et al. (2014).[9] Consider the color phi phenomenon. The display simply consists of a brief flash of a red circle on left and then a brief flash of a green circle on the right. When the temporal gap between the two flashes is small, we see a single "object" in motion that changes color midway through its trajectory. When the temporal gap is large, we see two separate flashes.

What's needed is an organizing principle that specifies when either a single

[7]This may have been what Pylyshyn was trying to do in the first place when he posited FINSTs as a way of individuating objects independent of the sensory content of an object file.

[8]It's possible, of course, for a visual system to construct two distinct object files for the same object. This would be an example of a "Frege" case, where two distinct representations refer to the same thing.

[9]Indeed, we find a general explanatory convergence across all these views. Brovold and Grush(2012), for example, see individuation by Gestalt principles (what they call "gobject" individuation) as playing the same role as Pylyshyn's FINSTs: "The perceptual system isolates these gobjects, assigns an object file to them, binds features (e.g., pitch, volume, timbre), and tracks them over time..." (p. 17).

object file is opened (and how it is maintained), or when two object files are opened (and how they are maintained). For the color-phi case, we need an organizing principle that uses sensory information—e.g., the shape, color, and spatiotemporal properties—and generates a decision about whether it should open one object file (and re-assign the object a new color part way through the trajectory), or whether it should open two separate files (and assign each a different color). In other words, what is required is an abstract psychological structure that flexibly integrates sensory information variety of visual (and non-visual) sources for the purpose of individuating objects—i.e. a structure that orchestrates the various perceptual cues along the dimension of "objecthood" or "thing" (i.e., as opposed to other visual properties such as color or shape, and other ontological categories, such as "liquid stuff" or "space").

Pylyshyn correctly identified the problem facing descriptivist theories of object individuation. And he may be correct that the visual system possess a primitive visual indexing mechanism. But object individuation involves more than tracking the spatiotemporal paths of objects.

## 6.5 Are Objects Perceived?

I argue above that one of the guiding theoretical themes in the vision sciences is the project to identify and characterize the underlying organizing principles of the visual system. We have principles that assign depth from shading. We have principles that group elements in a scene into lines. In the previous section, I

articulated the need for *abstract* organizing principles that explain object perception, and in particular the representation of object persistence.

### 6.5.1 Spelke Arguments

The question arises of whether we should think of the organizing principles governing object individuation as genuinely perceptual. Recall from Chapter 4 that Spelke holds that object representation is best understood as a cognitive phenomenon. Spelke offers two general arguments for why we *conceive*, as opposed to *perceive*, objects. The first argument concerns the abstract nature of the properties that constrain object representations (cohesion, boundedness, rigidity, spatiotemporal contiguity):

> Each of these properties is abstract: It cannot be seen or smelled or touched. These properties can be known, however, because each constrains how objects can be arranged and how they move (p. 226).

It's unclear from Spelke's discussion why she thinks cohesion, etc. cannot be perceived. She may be appealing to introspection, or a naive model of perception as mere sensory registration. At any rate, let us call this the *argument from abstract properties*.

The second argument concerns the observation that object representation involves parsing the world into distinct units:

> Human perceptual systems appear to analyze arrays of physical energy so as to bring knowledge of a continuous layout of surfaces in a state of con-

> tinuous change. We perceive the layout and its motions, deformations, and ruptures. This continuous layout contains no spatially bounded "things" and no temporally bounded "events": Perceptual systems do not package the world into units. The organization of the perceived world into units may be a central task of human systems of thought. (Spelke, 1988, p. 229)

This argument states that perception doesn't impose any ontological constraints upon the world. The *perceptual* difference, say, between the ocean and the beach is just a difference in color and visual texture. Only cognition can categorize the water and the sand as distinct things. Let us call this the *individuation argument*.

I will begin with the *argument from abstract properties*. The notion that we cannot see abstract properties is often motivated by paradigmatic instances of abstract properties. I cannot perceive the current market value of a Mickey Mantle rookie card because that property is realized by a host of non-local, relational facts. To a first approximation, the value of a Mickey Mantle rookie card depends on how strongly people covet the card. And it's very hard to comprehend how someone could perceive that property.

The issue of whether we perceive abstract properties is made more difficult due to the fact that there is no standard account of the distinction between abstract and concrete. Let us stipulate that a property is abstract if and only if it is not located in space or time. Conversely, a property is concrete if and only if it is located in space or time. Relations are typically thought to be abstract in this sense. If

the *relation* between the beliefs and desires of baseball card collectors and a Mickey Mantle card is what instantiates its market value, then that property is abstract. Likewise, cohesion, boundedness, rigidity, and spatiotemporal contiguity all seem to be relational properties, as they concern the relations between parts of objects.

At an intuitive level, I understand the motivation behind this line of thinking. It's hard to conceptualize how we "see" relations, especially if on thinks of visual perception as picture-like. Although elements in pictures have relations, they don't represent these relations, as such.

However, from a more theoretical standpoint, it's unclear why we should think that the visual system is unable to track relational properties. Our best theories of color perception, for example, treat perceived colors, not as monadic "simples" (as many philosophers think) but as relational properties—i.e., locations in color spaces (Clark, 2000). Indeed, when we reflect upon the earlier example of contrast constancy, the visual system must keep track of a variety of spatial and luminance relations in order to generate the effect. There's simply no other way to explain contextual effects of this sort.

Note, further, that while cohesion or boundedness relations might be abstract, the *relata* are plausibly concrete and spatially local. When one is perceiving a "medium-sized" object, the spatial relationships are often entirely within one's field of view. This is not true of the Mickey Mantle card example. So perhaps we can only perceive relations where the relata are within our view. But granting this much, I see no way to account for our uncontroversial perceptual abilities without positing perceptual structures that keep track of relations. So let us now turn to

the *individuation argument*.

It's often claimed that perception is non-conceptual (Block, 2014; Burge, 2010; Bermúdez, 1995, 2003; Dretske, 1999; Peacocke and Christopher, 1992; Peacocke, 1998). In addition to this negative characterization, we find a variety of attempts to provide a positive characterization of the format of perceptual content. Dretske (1999) claims that perceptual content is analogue, as opposed to digital. Analog representations are continuous representations, whereas conceptual/digital representations impose particular category boundaries. Similarly, Fodor (2003) suggests that non-conceptual content should be understood as iconic or picture-like representations, as opposed to discursive representations. The main difference between the two formats is that discursive representations are compositional—i.e., they have canonical parts that can be recombined—whereas iconic representations are not. As Fodor (2003) puts it, when decomposing iconic representations into parts, "there is no distinction between their *canonical* parts and their *mere* parts" (Fodor, 2003, p. 108). At the heart of these proposals lies the idea that perception lacks the cognitive machinery to divide the world into distinct kinds of things. Only conceptual machinery is able to do that, and we only find concepts in cognitive systems. If this is correct, my earlier discussion concerning object individuation must be incorrect. The type of machinery that I claim is necessary to explain object individuation presupposed individuating items in the environment *as* objects (or, at least proto-objects).

### 6.5.2 Categorical Perception

While there may be some merit to these proposals, they are misguided in claiming that perception doesn't individuate aspects of the visual scene.[10] First, it would be difficult to make sense of uncontroversial cases of organizing principles without thinking that they package, at least in a primitive form, sensory information. Gestalt grouping procedures single out and group particular contours as belonging to the same line. Of course, Gestalt grouping principles don't involve high-level categorization, but they do involve individuating aspects of the proximal stimulus.

Second, there's now good evidence for what is known as categorical perception. First noticed in speech perception, categorical perception involves the classifying or "chunking" of sensory information. For example, the difference between the consonants /ba/ and /pa/ lies along the dimension of "voice onset time" between 0 and 70ms (i.e., the time between the start of the consonant sound and the vibration of the vocal chords, or the "voicing"). Despite voice onset time being a continuous variable, the perception of the different phonemes is marked by a relatively sharp boundaries. For voiced consonants with an onset time of 35ms or below, the consonant is heard as /ba/, and for voice onset times above 35ms, the consonant is heard as /pa/ (Harnad, 2003). While once thought to be limited to linguistic perception, categorical perception actually exists in a number of different domains (color vision Goldstone and Hendrickson 2010; music perception Patel 2008; face perception Bestelmeyer et al. 2008; biological motion perception Sweeny et al. 2012), and

[10]Note: Fodor (2003) is merely attempting to explicate non-conceptual content. He doesn't claim that perception is restricted to non-conceptual content.

in a variety of different species (non-human primates Ramus et al. 2008; crickets Wyttenbach et al. 1996; and birds Nelson and Marler 1989).

Psychologists operationally define categorical perception as involving two phenomena. First, stimuli that range along a physical continuum are "perceived as belonging to distinct categories, rather than gradually changing from one category to another." Second, stimuli "of a given degree of physical difference are much easier to discriminate if they straddle a category boundary than if they fall within the same category" (Patel, 2008, p. 37). Evidence that the first criterion is met generally derives from subjective reports: red appears as a distinct color from orange, a /ba/ sounds distinct from a /da/, and so on. Evidence for the second criterion derives from careful psychophysical measurements that determine either the "just noticeable difference" (the smallest difference in stimulus values that individuals can reliably discriminate) or the "point of subjective" equality (the maximum difference between stimulus values where discrimination is at chance).

The second criterion for categorical perception gives us good reason to think that instances of categorical perception are genuinely perceptual. Finding that there is finer discrimination for stimuli pairs that straddle a category boundary than there is for stimuli pairs within a category suggests that the perceptual processing actually chunks the sensory signal, as opposed to post-perceptual processes applying labels to locations on an undifferentiated sensory dimension. That is, the processes responsible for the finer grained discrimination across category boundaries seems to explain the categorization. The hypothesis that perceptual structures are responsible for defining the categories, furthermore, explains why physically similar acoustic

signals (/ba/ and /da/) can sound so strikingly different.

The fact that we find uncontroversial evidence of perceptual categorization puts considerable pressure on the "argument from object individuation"—the idea that visual representation lacks that machinery to individuate perceptual entities. The existence of categorical perception doesn't entail that we visually represent object. But it does deflate the intuition that vision must be wholly iconic.

Given this result, a natural question is: are there any positive reasons for thinking that objects are represented? Here is where thinking about object representations as object files is helpful.

### 6.5.3 Why Objects are Perceived

As I pointed out above, object individuation seems to require flexibly integrating cues from a variety of sources in the retinal projection.[11] What we often find is an interdependence between object individuation parameters and other paradigmatic visual parameters. The idea, here, is that this sort of independence is good evidence that object individuation occurs within the visual system itself, as opposed to a post-perceptual judgment.

Consider apparent motion. Two dots in different locations will only appear as one dot in motion if the spacing is not too large and the offset/onset interval is relatively small. Subtle differences of the offset/onset interval determine whether we see one "object" (in motion) or two "objects" briefly flashed in quick succession.

[11] It also involves integrating sensory information from non-visual sources, but this complicates the picture, so I shall abstract away from multi-modal sensory integration.

Of course, anyone remotely familiar with visual psychology is aware of this fact. But what is noteworthy about this standard example is the reciprocal dependency between motion attribution and object file "management." The visual system imposes constraints on how cues for motion and cues for object individuation resolve conflicts and settle on an interpretation, and these constraints involve a subtle interdependence between the conditions under which objects are individuated and other paradigmatic visual processing.

We can see this interdependence when motion cues are manipulated. For instance, Ramachandran and Anstis (1986) report that apparent motion can be induced for displays that are not normally viewed as apparent motion when additional motion cues are provided. The left side of Figure 6.5 shows a dot on the left and a square on the right. When the dot is extinguished, observers simply report that the dot disappeared. However, when normal apparent motion displays are stacked above and below the dot/box display, the dot appears to move behind the occluding box. That is, adding congruent apparent motion is enough to generate the illusion of motion for the dot/box display. What's likely going on in this case is that the normal apparent motion displays provide evidence that the middle dot moved as well. Because object individuation seems to depend on motion perception and vice versa, this gives us reason to think that object perception is perceptual.

Also, recall Papenmeier et al.'s (2014) result, where spatiotemporal and chromatic properties contribute towards tracking an object. In this case, when spatiotemporal cues are ambiguous or otherwise less reliable, they receive less weight

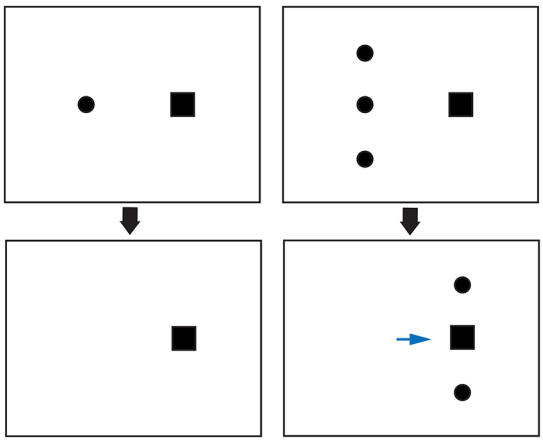

*Figure 6.5: Contextual cues for apparent motion. The top panels depict the initial display; the bottom panels depict the end state of the display. In the left panels, when the dot extinguishes, observers report not apparent motion. In the right panels, the contextual apparent motion biases one to see the middle dot move and disappear behind the box. (Original drawing based on descriptions in (Ramachandran and Anstis, 1986))*

in the final computation of an object's trajectory. So we not only have sensitivity to changes in the stimulus, but also subtle interaction effects between the *reliability* of visual information and how that information is used. We typically don't find post-perceptual sensitivity to the reliability of a cue.

As it turns out, this kind of subtle interdependence between visual properties is a general feature of visual processing (Churchland et al., 1994). Indeed, we find consilience between the interdependence of psychophysical variables and how brain areas are individuated. Connectivity is an important criteria for distinguishing different brain areas. Borders can be drawn between different cortical areas by measuring the amount of connectivity between any two regions. The boarders are defined as low contours where connectivity density is at its lowest (Hilgetag et al., 2000; Markov et al., 2013).

Furthermore, when we look at the the boundaries defined by changes in con-

nectivity density, we find that they line up quite well with boundaries defined by retinotopic maps. So it seems that within visual area connectivity is more dense than connectivity that crosses this border; therefore, one expects greater interdependence between psychophysical variables when the belong to the same system.

## 6.6 Conclusion

I have argued in this chapter that organizing principles are a fundamental posit in the vision sciences, and I have sketched the role that they play in explanations of paradigmatic visual processes. I also argued that we need to posit organizing principles to explain object individuation. In the language of object files, organizing principles determine which contents get bound into an object file and govern the ways in which object files are updated.

One thing worth emphasizing is the abstract nature of object files. We've been thinking of object files as consisting of a set of principles that specify when object files are created, and how they are maintained. I have proposed that it's in virtue of these principles that we visually represent objects. That is, object files organize the sensory input along the dimension of *objecthood*. If this is correct, we have already established that the visual system represents high-level abstract properties. Therefore, there's a sense in which Spelke is correct that objects are *conceived*. In order to perceive objects, as such, we need (in her words) an "object concept" that governs how our minds segment and individuate objects into coherent wholes. Where Spelke errs is thinking that this form of abstract representation is

incompatible with visual processing.

# Chapter 7: High-Level Visual Schemata

In the previous chapter, I argued that we should think of the visual system as possessing a variety of different psychological structures that organize the sensory input. These organizing principles are responsible for encoding features of our environment that range from low-level contour processing to high-level object perception. In this chapter, I extend this idea to cases of causality, animacy and facial categories of emotion, race, and gender. I argue that the visual system possesses high-level representational structures that are organized along abstract dimensions.

I conclude this chapter by proposing we think of these representations as Kantian schemata. Schemata are "intermediary" representations that straddle the sensory and the conceptual domains. They form a bridge between the endlessly changing representations of the sensory domain and our stable and enduring conceptual schemes. To distinguish these high-level organizing principles from their low-level brethren I shall adopt Kant's terminology for this distinct class of representations for the remainder of this chapter.

## 7.1 Perception of Causality and Animacy

This section deals with two distinct, but related, areas of research: the perception of causality and the perception of animacy. Although causation and animacy are distinct metaphysical categories, in the sense that understanding the nature of causation is unlikely to reveal much about the nature of animate objects (and vice versa), work on how humans cognize these aspects of our world are often linked as research programs. For instance, research on the perception of causality and animacy both involve relatively sparse two-dimensional displays of moving objects that give rise to seemingly high-level percepts. Furthermore, theorists appeal to similar considerations when defending the idea that humans perceptually represent these abstract properties. Thus, researchers who are theoretically and methodologically invested in one area are theoretically and methodologically invested in the other. If for no other reason than exegetical convenience, I shall discuss these two areas of research in tandem.

### 7.1.1 Causal and Animacy Displays

Michotte's (1963) work on causal perception is the starting point for contemporary research on the topic of causal perception. He found that remarkably simple displays of two-dimensional shapes moving in patterns characteristic of actual mechanical interactions elicit reports of "impressions of causality."

The most familiar example of a causal impression is what Michotte referred to as a "launching" event. A typical launching display is a follows. Starting on the

left, object A moves horizontally at a constant velocity towards towards object B until it is directly adjacent to object B, at which point A stops, and B continues along the same trajectory. When people view such a display they typically report *seeing* A cause B to move. Naturally, in order to "perceive" (or judge) that A caused B to move, the spatiotemporal properties of A and B must approximate the spatiotemporal properties of an actual collision of objects. For example, if there is too much of a delay between the termination of A's movement and the initiation of B's movement, subjects will report two independent events—A moves, then B moves. (White, 2012). Recent research confirms Michotte's primary findings that participants *report* directly seeing causal interaction, and that these reports depend on subtle changes to spatiotemporal dynamics of two-dimensional shapes (Gordon et al., 1990; Schlottmann et al., 2006; Scholl and Tremoulet, 2000; Scholl and Nakayama, 2002; White and Milne, 1997; White, 2012). The primary question is whether we have reason to think that causality is represented when participants view a "causal" display, or whether a more deflationary explanation is in order, on where the representation of causation is entirely post-perceptual. (see Rips, 2011).

The perception of animacy concerns the visual analysis of animate behavior. One can think of the distinction between animate and non-animate "objects" in the following way. Non-animate objects tend not to move. But when they do, they tend to move in accordance with Newtonian mechanics in rather straightforward ways. Unsupported non-animate objects fall downwards until they collide with a surface. Non-animate objects at rest remain at rest unless they are moved by some other object. Animate objects, however, can move on their own accord, resist gravity,

and change course without any external causal-mechanical interaction with another object (Tremoulet and Feldman, 2000).

Work on animacy perception also has its roots in mid-20$^{th}$ psychological research. Heider and Simmel (1944) created an animation consisting of three shapes (one large triangle and two smaller shapes) moving around and into a rectangular box with an opening on one side. In describing what they observed, participants tended to use intentionally-laden locutions, such as "the circle goes out of the opening and *joins* the smaller triangle" and "and when the larger triangle comes out of the rectangle and *approaches* them" [emphasis mine] (Heider and Simmel, 1944, p. 246). While Heider and Simmel's experimental design lacks some of the controls that modern psychology requires, more recent experimental designs support their initial results that observers naturally describe the elements in two-dimensional displays as having certain goals and intentions.

### 7.1.2 Scholl's Argument from Modularity

The central issue of this section is whether there is reason to think that humans perceptually represent causality and animacy. Scholl and colleagues offer a number of considerations in support of the thesis that the visual system represents causation and animacy. The primary argument offered (though not the only one) is that the underlying mechanisms supporting these capacities bear the hallmarks of modularity. Scholl and Tremoulet (2000), for instance, claim that causal and animacy perception is (a) fast, (b) domain specific (in that the percepts are about causa-

tion and animacy), (c) automatic (i.e., the impressions of causality and animacy are irresistible), and (d) encapsulated (e.g., knowing that one is looking at mere two-dimensional shapes doesn't extinguish the impression of causality or animacy).

While I think that causal and animacy perception by and large possess these sort of properties, the argument from modularity leaves many questions unanswered. The first problem concerns the claim of domain specificity. Scholl and Tremoulet (2000) seem to think that this claim falls immediately out of their observations: "They are domain-specific by definition, in that they result in specific causal and intentional interpretations (i.e., they give rise to only a few qualitatively separate types of percepts)" (p. 306). Notice, however, that the issue of domain specificity typically concerns the content of the information processed by a module. If one knows that a module's domain is restricted to animacy, then one can conclude that the content of outputs of the module will be about animacy. However, it's not clear what grounds the domain specificity claim.

One possibility is that Scholl and company think that one's phenomenology provides good evidence that the visual modules in question output causal and animacy outputs. Scholl and Gao (2013), for example, suggest that one can *see* the animacy in certain displays:

> In the first place, it is worth noting explicitly that the rich phenomenology of perceived animacy is consistent with an interpretation in terms of visual processing. Indeed, the phenomenology elicited by such displays is surely the driving force behind this research program as a whole. These

> phenomena often operate as fantastic demonstrations: observers simply *see animacy and intentionality when viewing the displays...*[emphasis mine] (p. 207)

Gao et al.'s (2010) "wolfpack" display provides a compelling example of the sort of "rich phenomenology" referred to in this passage. This display consists of randomly moving "darts" (the wolves), and a number of discs (one of which is the "sheep"). To induce the sense of animacy, the darts can be programed to continuously orient towards one particular disc. When this occurs, the "sheep" becomes immediately salient (it "pops out" from the other distractor shapes), and one gets the sense that the wolves are chasing it.

Visual saliency, however, is generally understood as arising from attention. Here, a low-level theorist might argue that the darts direct attentional resources towards a particular circle, facilitating the processing of its features. Plausibly, the additional processing that the target circle receives would be enough to generate a pop-out effect and alter one's phenomenology of the low-level features.[1] Since Scholl and colleagues don't rule out the possibility that the distinctive phenomenology is due to attention, one must look to something else to secure the claim that we visually represent animacy.

What about the other properties, such as the speed, automaticity, and encapsulation of causal and animacy perception? If we have evidence of a animacy module does this give one reason to think that causation and animacy are visually

[1] Recall my discussion in Chapter 3 of the Carrasco's (2004; 2011) work that shows that attention can alter visual appearances.

represented? The problem is that the low-level theorist can agree that spatiotemporal information is processed quickly, automatically, and independently of higher cognitive systems. But this system doesn't output animacy-level representations. If Scholl and colleagues assume that animacy *percepts* are quick, automatic, and encapsulated, then they have simply begged the question. So the question is, given the data, must one posit fast, automatic, and encapsulated visual modules that output causation and animacy percepts?

Rips (2011) argues that on balance theorists do not need to posit animacy percepts, and offers an alternative explanation for the data. He suggests that when an observer is watching a causal display, visual processes produce a low-level percept. The spatiotemporal properties of this percept are compared with various schema stored in long-term memory. When the visual content matches a "causal schema," a post-perceptual judgment is issued. Note that Rips' alternative model supposes that the schemata are stored in higher level cognitive systems and, plausibly, would be susceptible all kinds of cognitive influences. Scholl and colleagues would likely respond by arguing that this latter claim is undermined by the available evidence. For example, the fact that causal impressions persist despite knowing that one is looking at a computer generated two-dimensional display suggests (at least provisionally) that causal judgments are not part of central cognition.

However, one might think that causal and animacy judgments are performed by dedicated, post-perceptual modules. These modules contain their own long-term memory stores, but are isolated from other forms of cognitive influences. If this were the case, one might expect causal and animacy judgements to exhibit all

the hallmarks of modularity without being perceptual in nature. The basic point here is that evidence of modular processing doesn't entail evidence of perceptual processing. Modularity might be ubiquitous throughout the mind, as a number of theorists propose (Barrett and Kurzban, 2006; Carruthers, 2006; Pinker, 1997; Sperber, 2002).

### 7.1.3 Argument from Categorical Perception

In a number of articles, Scholl and colleagues appeal to the fact that causal reports are sensitive to subtle visual details: "We suggested above that a hallmark feature of perception (vs. cognition) is its strict dependence on subtle visual display details" (Scholl and Gao, 2013, p. 209). But note that the mere fact that a perceptual report is highly sensitive to visual details doesn't show that the report reflects perceptual processing. An art appraiser's judgment about whether a work is a forgery or a genuine Picasso might be highly sensitive to subtle visual details, but this doesn't demonstrate that she visually represents these high-level properties (i.e., *being a forgery* and *being a genuine Picasso*).[2]

Butterfill (2009) and Carruthers (2015) advance a similar (though more promising) strategy, arguing that causal and animacy perception involve categorical perception. If it can be shown that perceptual processes are organized categorically along the dimensions of causation and animacy, then one would have evidence of causal and animacy schemata. However, despite the promise of this approach, neither

[2]Showing subtle interaction effects between causal determinations and low-level visual processes is a different story, and I shall discuss Scholl's work on this point below.

Butterfill nor Carruthers provide adequate evidence to conclude causal or animacy perception are instances of categorical perception.

Butterfill appeals to Michotte's original data on the effects of manipulating the temporal delay in the launching displays to argue that humans visually represent causal relations. He argues that Michotte's data show that perceivers see causal relations categorically. The main effect is illustrated in Figure 7.1. For delays between 14 and 63 ms, participants all viewed the display as a causal launching. After 63ms, reports of a causal launching begin to drop precipitously. Now, recall that categorical perception involves both categorical reports *and* better discrimination performance across category boundaries than within category boundaries. The data here support the first criterion, namely, that observers categorize stimuli that range along a continuum; one can reliably predict how an individual will perceive the display from the temporal delay.

However, Butterfill also concludes that observers are unable to discriminate between launching displays with a delay of 28ms when this interval lands within either category (either the "launching" category of the "two separate events" category):

> Subjects would not normally be able to distinguish stimuli between trials when the stimuli differ only in that the gap between movements is 29 ms longer on one of them; but they can easily discriminate stimuli in the special case where 28 ms make the difference between the experi-

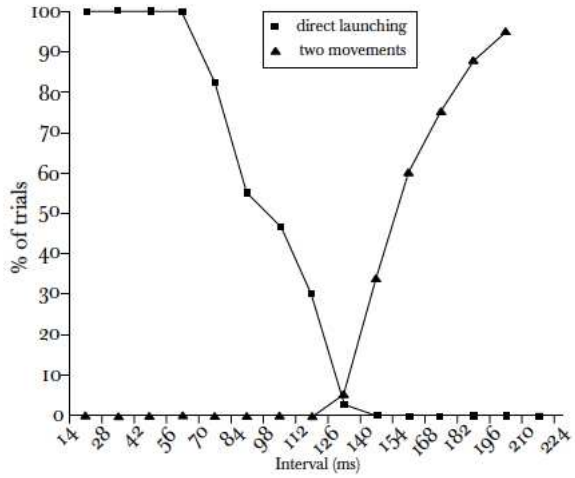


*Figure 7.1: A graph constructed from Butterfill (2009) using Michotte's (1963) data. This chart plots causal vs. non-causal judgments ("direct launchings vs. two movements") as a function of the interval between the first object stopping and second object initiating movement in the launching display. Reprinted with permission.*

> ence characteristic of launching and the experience characteristic of two movements." (p. 417)

Butterfill's thinking is that people typically don't notice small differences in temporal delays (and 28ms is really quite small), but when 28ms straddles the boundary between launching and non-launching events, that difference is readily perceived.

While I think it's plausible that casual perception is categorical in this sense, this is not what the data show. The task that Michotte's participants performed was a categorization task. They were simply asked to identify whether a given display was a launching or non-launching event. The fact that people categorize stimuli at a particular point in the continuum is consistent with perceptual discrimination being identical across the continuum. For all Butterfill has shown, participants might be able to discriminate between two intra-category displays where the temporal delays differ by 28ms or less.

In order to show that across boundary discrimination performance is better than within-category discrimination, participants would need to perform, for example, a pair-wise discrimination task for all (or all relevant) stimulus values along the continuum. One could then construct a similarity ranking for each pair, and observe whether within-category pairs (of a given distance apart on the continuum) are more similar than those that straddle the category boundary. If across-category discrimination is better than within-category discrimination, then can conclude that causal perception is categorical. But from the data Butterfill provides, we can't conclude this. As far as I am aware no experiments have been performed to address this question.

Arguing a related point, Carruthers (2015) claims that the perception of animacy is a case of categorical perception. Carruthers takes aim at the possibility that animacy perception is non-conceptual and therefore perceptual. If animacy perception is nonconceptual, then one should expect that different animacy "percepts" should differ by degree as opposed to strict boundaries. But if animacy perception exhibits the hallmarks of categorical perception, then this would suggest it is not nonconceptual. Independent of this line of argument, if Carruthers is correct that animacy perception is categorical (in the technical sense), then this provides evidence that a high-level schema underlies animacy perception.

Carruthers appeals to Gao et al.'s (2010) work using the "wolfpack" displays. He notes that in some displays, the "wolves" will alternate between two targets. That is, the darts will first point to one disc, then another, then back to the first and so on. When the wolves alternate between targets in this way, participants

don't report a partial chasing, but rather a distinct switch of targets, even when it could be ambiguous which of the discs was the target. Carruthers concludes: "This suggests that one's more-or-less vivid impression of animacy is really a categorical perception of animacy in which one has more-or-less *confidence*...(p. 501)." However, Carruthers does not provide additional data showing that participants more finely discriminate between stimuli that straddle the boundary between stalking and non-stalking, nor am I aware of any experiments that demonstrate this effect. But without evidence for a difference in discrimination, he hasn't shown that the wolfpack effect is a case of categorical perception.[3]

So far I have been discussing causal and animacy perception in tandem. I now address these two phenomena separately, as the evidence is stronger for causation than it is for animacy.

### 7.1.4 Evidence for Causal Schemata

Rolfs et al. (2013) find evidence of spatially specific adaptation aftereffects for "launchings." This is a nice experiment for our purposes for two reasons. First, the aftereffect profiles provide evidence of categorical encoding, suggesting that our brains (at some point in the processing hierarchy) deploy a causal schema. Second, because the aftereffects are spatially specific—i.e., an aftereffect can only be generated when the "adaptor" and the test stimulus are presented in the same visual hemi-field—we have reason to think that the locus of adaptation is anchored within

[3]Recall from the previous chapter that the within/across category discrimination differences is what justifies the claim that the categories are perceptual.

a visual coordinate scheme, and therefore the causal content is visually encoded.[4]

The experiment relies on the finding that when the two objects (or discs) overlap too much at the "point of contact," participants no longer see the first object colliding with the second. For the first phase of the experiment, participants view two launching displays (one on either side of a fixation point) that vary in overlap from trial to trial. Their task is to identify whether the target display (either the display on the left or the right, depending on the trial) is causal or non-causal. This data provides the participants' baseline performance, and is used to plot the "proportion of non-causal" judgments as a function of the amount of overlap. If participants adapt to causal events, then one expects the curve to shift, such that stimuli that originally appeared causal at baseline will subsequently appear ambiguous post-adaptation (and stimuli that originally appeared ambiguous at baseline will subsequently appear non-causal).

The second phase of the experiment involves participants adapting to a causal stimulus on either the right or left side (depending on which side they performed the task in the first phase). The adapting stimulus consists of a stream of 320 "launching" events (without any overlap). Importantly, the direction of motion rotates randomly throughout the adapting period.[5] The third phase of the experiment is identical to first, except half of the participants are asked to perform the "causal–non-causal" task on the stimuli in the opposite hemi-field. This allows ex-

[4]As I argue in Chapter 5, spatial specificity is suggestive, but not probative evidence of visual processing.

[5]The stream consisted of a sequence of pairs of launching events, with the first event of the pair having a random direction of motion and the second having the opposite direction of motion. One pair of events would look like a "ping pong" event, where A hits B, B moves to edge of screen then moves back towards A.

perimenters to determine whether the adaptation effect is spatially specific (though not retinotopically specific). If the adaptation is not spatially specific, then the aftereffect should transfer across hemi-fields.

Rolfs et al. (2013) found that adapting to a stream of causal interactions, indeed, shifts participants' perceptual reports. That is, stimuli that were once perceived as ambiguous are subsequently perceived as non-causal post-adaptation. However, this was the case only when the adaptor stimulus and the test stimulus were located in the same hemi-field.

What does this result tell us? One can rule out the possibility of adaptation to direction-of-motion as a possible confound, because the adapting stimuli constantly rotated during the adapting period. However, because the adapting stimuli all shared the same spatiotemporal contiguity, adaptation to this property could potentially explain the results.

Rolfs and colleagues considered this possibility and designed a control experiment. They used the same basic experiment setup. However, instead of having participants adapt to causal interactions, they had them adapt to similar (ping-pong-like) non-causal "slip" events. They reason that if the adaptation effects in the first experiment are driven by low-level details (spatiotemporal contiguity, timing, etc.) of the scene, then the non-causal "slip" adaptors will generate comparable aftereffects. This is not what they found.

Instead Rolfs et al. (2013) report that "adaptation to slip stimuli had little or no effect on observers' perceptual reports" (p. 252). That is, adapting to slip stimuli does not shift how people perceive/judge the subsequently viewed stimuli.

These authors conclude that the aftereffects generated by the causal stimuli in the first experiment are not driven by low-level visual details, but rather by the repeated *causal* events.

This result strongly suggests that the visual system embodies what I have been describing as a mechanical-causal schema. The adaptation effects are driven by a narrow, but disjointed, collection of visual parameters. In order to generate the aftereffect, one needs a very particular sort of spatiotemporal coincidence, namely spatial contiguity (overlapping coincidence won't do). Yet, the aftereffect does not depend on any specific pattern of stimulation, as the stream of adapting stimuli continuously changed their direction of motion.

It's important to stress that the slip events do not generate any aftereffect. Not only does adaptation to slip events fail to alter how subsequent causal events are perceived, they fail to alter how identical slip events are subsequently perceived. This is important because it shows that not all complex motions induce aftereffects.

This raises the following question: Why do we find adaptation to *this* specific (but disjoint) set of visual parameters, and not others? Recall my discussion of Gestalt grouping processes in Section 6.3 of the previous chapter, where I asked, "Why do we see this particular configuration of lines and not another?" My suggestion was because we possess a set of organizing principles (or a schema) that structures the sensory input according the proximity and good continuation of the elements. We can provide an analogous answer to the question of causal adaptation: We adapt to *this* specific set of visual parameters because we possess a set of organizing principles (or a schema) that organizes the sensory input according to a

particular spatiotemporal pattern.

Now that there's good reason for thinking that an abstract causal schema underlies the psychophysical profiles in Rolf et al.'s (2013), I need to address the following question: Is the schema really perceptual? In my discussion of Block (2014), I argued that the mere fact that one finds adaptation doesn't show that the adapted attribute is perceptual in nature. Likewise, the mere fact that one finds a reason to posit an abstract schema to explain the causal adaptation does not mean that it is perceptual in nature. So in order to conclude that we perceptually represent causal interactions, I need independent reasons for thinking that the schema is perceptual.

As I argued in Chapter 6, one of the hallmarks of visual processing is an interdependence between variables. So evidence that causal judgements interact in subtle ways with other paradigmatic visual variables, will provide reason to think that the causal schema is perceptual in nature. And, indeed, Scholl and Nakayama (2002) find such evidence.

They show that non-causal slip events can be made to appear as causal events when the slip events are adjacent to unambiguous causal events (in much the same way as apparent motion can be generated with the appropriate visual context). For example, if object A completely overlaps object B prior to B initiating movement, A and B's movement will be judged as two separate events. But if observers see an unambiguous launching event situated just below a slip event, they will report the slip even as a causal event. They term this effect "causal capture." This suggests that perceptually representing the one event as a causal interaction biases the causal

attribution of the other.

Interestingly, when observers view the slip event combined with an unambiguous causal event, they underestimate the amount of overlap. Informally, Scholl and Nakayama noticed that for 100% overlap events (in the context of an unambiguous causal event) the discs do not appear to completely overlap. That is, it appears that object A does not completely cover B, but rather leaves a thin "crescent" of B uncovered. It seems then that a causal context not only biases causal judgments, but alters the appearance of the overlap.

To test this, Scholl and Nakayama (2004) presented participants with launching–slip hybrid displays with varying amounts of overlap. Participants were asked to adjust the amount of overlap on two comparator discs. This allowed them to measure to the amount that participants underestimated the overlap. In order to allow participants enough time to adjust the comparators to match the amount of overlap, the events were iterated. (A moves towards B, then B moves until it reaches the edge of the screen. B then moves back towards A, than A moves, and so on. This repeated until the participants submitted their response.)

Scholl and Nakayama (2004) found very little bias for 60% overlap. But as the overlap increases (80, 90, 100%), they found increasing divergence between perceived amount of overlap in the absence of contextual causal events and the perceived amount of overlap in the presence of contextual causal events. The presence of the unambiguous causal display causes observers to underestimate the amount of overlap.

What's remarkable about these results is how a launching event perceived in

one location of the scene alters reports of the amount of disc overlap in another location of the scene. One could argue that Scholl and Nakayama's (2002) initial results merely show that the addition of an adjacent causal display alters participants post-perceptual judgments. Perhaps the judgment that one event is causal biases judgments about the other event. But it's not obvious how this explanation is supposed to account for the underestimation of overlap. Remember, what is driving participants' reports is how the overlapping disc look. Their task was to make the comparator discs look the same as the amount of maximal overlap in the slip displays. So the effect, here, seems to be a visual effect.

One other possible explanation would be to argue that causal judgments penetrate visual processing, biasing how participants perceive the overlap. I have no objection to the possibility of cognitive penetration, but it's unclear what would motivate this view. Normally, cognitive penetration is posited when stored information is used to modulate visual processing. But in this case, all the relevant variables derive from the scene. On this hypothesis, the visual system would have to processes the low-level details of the unambiguous causal event to produce a non-perceptual judgment of causation. This judgement would then be fed back down to the visual system bias the how it interprets the slip event. This seems rather implausible.

According to the view that I am proposing, the visual system possesses an abstract causal schema that tracks how objects interact and produces a perceptual judgment of causation under certain circumstances. Because it's part of the visual system, the various cues are processed in parallel, such that visual judgments concerning various aspects of the scene need to (more or less) cohere with each other.

As such, where there is conflicting information, the conflicts need to be resolved. Hence, when additional contextual information provides evidence of a causal interaction, the visual system attempts to resolve this conflict by biasing the visual judgment of overlap. That is, there is plenty of motivation for thinking of Scholl and Nakayama's result is a visual phenomenon, and very little for thinking that it is a non-visual judgment.

### 7.1.5 Evidence for Animacy Schemata

The evidence for high-level animacy schemata is not as nearly as robust as the evidence for perceptual schemata for causality. However, Scholl's observations that animacy judgments are fast, automatic, and immune to conflicting background beliefs gives one reason to think that animacy content fixations are the result of a modular system, as opposed to higher cognitive judgments. The problem described earlier in this chapter was the lack of evidence to conclude that animacy "judgments" are perceptual in nature. If there's evidence that suggests that animacy is processed in visual areas of the brain, then this provides reason to thing animacy is represented in vision.

There are a few imaging studies that point to areas within the superior temporal sulcus as site for animate motion processing (Blakemore et al., 2003; Castelli et al., 2000; Gao et al., 2012; Schultz et al., 2004). This area is associated with a number of socially relevant processes, such as face perception, biological motion perception, as well as the ability to interpret attribute mental states on the basis of physical

and linguistic behavior (Deen et al., 2015). Blakemore et al. (2003) find that this area is particularly sensitive to "contingent" animate motion—i.e., motion that is contingent upon the movements of another "agent."

Following up on this result, Gao et al. (2012) ran an fMRI imaging experiment using a similar chasing display to the one used by Gao et al. (2010) (described above). Instead of randomly moving darts that oriented towards a "sheep," they programmed a single disc to track (but never reach) a particular disc mixed in with other distractor discs. All of the discs except for the "wolf" move in a random fashion. Like other displays, observers get a distinct impression of which shape is being tracked and which shape is doing the tracking.

The experimenters then created three variations of the standard display to control for potential confounds. The first variation, the "Wavering Wolf" (or Changing Intention) display, was similar to the standard set up, except the wolf's target stochastically alternates between two targets. As in the Gao et al. (2010) experiment, participants are able to identify the wolf, and report that it switches targets.

Because switching targets generates abrupt wolf movements (more so than in the standard display), it's possible that a potential fMRI signature would be the result of the sudden movements, as opposed to the switching of targets. Experimenters therefore created a second display, the "Phantom Chasing" display. This display is identical to the Wavering Wolf display, except for the fact that the wolf chases non-visible sheep. Unlike Wavering Wolf, participants do not report seeing a chasing behavior in the Phantom Chasing display. Gao and colleagues reason that if the abrupt movements are driving activation then one would expect the same activation

profile in the Phantom Chasing condition as in the Wavering Wolf condition.

Finally, they constructed a "Flashing" display, which is identical to the Phantom Chasing display, except that each disc had a 10% chance of briefly disappearing. The aim of this display was to determine whether fMRI activity is driven by attention. It is well known that flashing recruits attentional resources. So if attention is driving patterns of brain activation, then one should see similar results for the Flashing, Phantom, and Wavering Wolf conditions.

Gao et al.'s (2012) results largely confirm earlier work that shows regions of the superior parietal lobe is involved in social perception. Their main finding iss that the Wavering Wolf display generates greater activation in the posterior superior temporal sulcus (pSTS) than for the Phantom Chasing and the Flashing displays.[6] They hypothesize that the pSTS is responsible for encoding changes in intentions, since the Wavering Wolf display generates greater activation than the single intention display. Here they make an important point that merits attention. Most fMRI animacy studies use basic animacy cues to invoke an impression of animacy, such as self-propulsion, and abrupt changes in speed and course. In these displays, manipulating animacy always involve directly manipulating low-level cues. As such, one might remain skeptical about whether pSTS activation actually reflects animacy representation, as opposed to the representation of a special class of spatiotemporal properties (those movements without environmental causes).

What's interesting about the Wavering Wolf display is that it seems to involve

[6]Gao et al (2012) did not find preferential activation for the single intention display, contrary to what earlier findings would suggest.

the application of a principle of rationality: "According to this principle, agents will tend to choose actions that achieve their desires most efficiently, given their beliefs about the world" (Gao et al., 2012). In the single intention display, the wolf "picks out" the sheep because it takes the an efficient (rational) route towards that object. Semi-formally, we might say that $\text{disc}_x$ chases $\text{disc}_y$ when $\text{disc}_x$'s path doesn't stray too far from the optimal path between $\text{disc}_x$ and $\text{disc}_y$. And, moreover, it applies this rule in a flexible manner. The Wavering Wolf display, however, routinely violates this principle. Every few seconds the wolf fails to take an efficient route towards a disc, at which point the visual system hunts for other efficient disc–route pairings. When faced with evidence that is contrary to the rationality principle, the visual system assigns a new goal to the wolf.

This line of evidence provides reason to think that the visual system represents agency, though the conclusions here are fairly speculative. However, it's plausible that one would find categorical perception of animacy displays. For instance, the animacy of chasing behavior can be easily modulated by altering the efficiency of the wolf's pursuit path by programing the wolf to randomly select a path with in a certain specified window. The larger the window, the harder it is to identify the wolf/sheep. Gao et al. (2009) find a fairly sharp drop in performance between 30° and 60° windows.[7] One could present two displays simultaneously and ask participants to indicate if they are the same or different, where the only possible difference is the efficiency of the wolf's path. If discrimination is better across cate-

[7]Gao et al. (2009) refer to this measure as the "subtly of chasing." The measurement in degrees refers to range of freedom on each side of the most direct path. Hence, a 30° subtly of chasing will produce a 60° range of possible paths.

gory boundaries, then this would suggest that animacy is categorically represented in perception.

Furthermore, one could look to see if manipulating the "existence" of animacy in a display alters other visual features. For example, does the presence or absence of animacy alter speed-of-motion or direction-of-motion perceptual performance? If it does, this would suggest that the animacy schema is embedded within the visual system itself. Again, it seems plausible that one would find this sort of evidence. Although the case of animacy perception could be stronger that it currently is, this is an active area of research. Importantly, however, a framework is in place to address these issues.

## 7.2 Representing Facial Categories

In this section I argue that humans visually represent facial categories, in particular emotional, gender, and racial categories. The burden of proof for positing high-level visual content should be familiar by now: I will show that some aspect of visual processing exhibits a categorical structure.

### 7.2.1 Emotional Categories

I will briefly re-examine the adaptation literature discussed in Chapter 5. Recall that one finds adaptation effects in response to facial emotions. After adapting to an angry face, participants will identify a neutral face as being frightened. The first question to address is whether the effects are categorical. As Block notes, one

reason is that that adaptation to faces that differ with respect to their low-level properties will nonetheless induce the same aftereffect profile, as long as they both have the same emotional content (see Butler et al., 2008). The idea here is that if the low-level details are different, then one would not expect participants to perceive ambiguous faces as frighted after adapting to an angry face. Yet the adaptation effects remain, suggesting that the adaptation is categorical in nature. The finding that emotional aftereffects persist despite low-level differences has substantial support (see Webster, 2011; Campbell and Burke, 2009).

But do these results show that we perceptually represent facial emotions? On their own, these studies certainly suggest this hypothesis. Observers typically report that a face looks different post-adaptation. Although this fact doesn't settle the issue, I hesitate to entirely disregard introspective reports as a source of evidence. What we need then is independent support of the thesis that categorical facial adaptation is perceptual in nature.

As it turns out, the categorical structure of face perception as been well studied. De Gelder et al. (1997) ran a standard categorical perception experiment for emotional categories. These experimenters artificially constructed three separate continua from images of individuals with different expressions of emotion: the first continuum ranged from angry to sad, the second from happy to sad, and the third from angry to afraid. Participants categorized the stimuli, and, using this data, the experimenters define the borders between categories as the point where categorization is at chance.

De Gelder and colleagues also asked the participants to performed a discrim-

ination task using images from the continua. Three images are presented in short succession. The first two are always different from each other. The third image is identical to one of the previous two images. The task involves determining which of the first two images the third image matched. (Discrimination performance bottoms out when matching is at chance.) As de Gelder et al. expected, discrimination is better for images that straddle category borders than for images within a category. Indeed, peak discrimination performance is centered over the previously identified category boundaries. Beale and Keil (1995) and Bimler and Kirkland (2001) provide additional support for the hypothesis that the visual system represents facial categories.

Because one finds a sharpening of perceptual discrimination at the category borders, there is reason to think that the categories have a perceptual basis. This suggests facial schema for emotions. De Gelder's results, then, provide indirect support for the hypothesis that emotional adaptation is a perceptual phenomenon. From the categorical perception literature, there is good reason to think that the visual system contains a categorical representational structure.

The evidence of perceptual schemata for facial emotions also provides a possible explanation for the facial adaptation effects. The adaptation studies show a categorical structure. It seems as though adaptation to angry faces makes neutral faces appear afraid. A perceptual schema that encodes facial features along emotional dimensions could explain both the psychophysical results and the phenomenology.

Findings in neuroscience provide further converging evidence for this hypothe-

sis. It's well established that areas within the superior temporal sulcus are dedicated for processing faces (1997; 2010). Recent functional and anatomical mapping studies suggest that the STS contains hierarchically organized category-specific feature maps, many of which pertain to faces. Grill-Spector and Weiner (2014) propose that each of these maps contain category-specific representation of space. Further, some of these areas are thought to be responsible for both facial identity and facial emotion recognition (Deen et al., 2015; Calder and Young, 2005) For example, areas within the STS (among other visual areas) preferentially respond to facial emotions (Sato et al., 2004). So we know that higher level visual areas are specialized for encoding facial identity and emotion.

This evidence provides support for the hypothesis that we perceptually represent facial emotional categories. If this is right, then the visual system possesses schemata for emotions, and it's plausible that the same schemata are responsible for the adaptation effects found in Butler et al. (2008). That is, the same mechanisms responsible for encoding facial emotion are the ones that undergo adaptation, and generate the aftereffects. Indeed, Winston et al. (2004) show that repeated exposure to a particular facial expression generates the characteristic suppression effect in STS. We therefore have a number of converging lines of evidence supporting the hypothesis that we perceptually represent high-level facial emotion categories.

### 7.2.2 Gender and Race Schemata

What about other facial categories, such as gender or race? Webster et al. (2004) report adaptation to both race and gender. Participants adapt to a random sequence of male or female face (depending on the trial). As expected, subsequent gender-based reports are biased away from the adaptor. There was a considerable amount of physical differences across the faces, suggesting that low-level features are not driving the aftereffect. Webster et al. also report similar racial after effects for Japanese and Caucasian faces. Both gender and race adaptation have been widely reported (Webster, 2011; Tiddeman et al., 2001; Javadi and Wee, 2012; Little et al., 2005; Pond et al., 2013; Leopold et al., 2005).

Yet, it's possible that there are perhaps holistic structural features that can define the the categories. Male faces are generally broader than female faces, for example. Perhaps, then, participants are adapting to wholistic structural features, and not gender, per se. On the basis of Webster et al.'s (2004) results that the adaptation profile reveals a categorical structure.

However, Bestelmeyer et al. (2008) provide strong evidence against this hypothesis. They reasoned that if adaptation were operating on structural features, then adaptation affects should be the same when they differ in the same respects. Their experiment relies on constructing sample faces using a prototype-based transformation algorithm. This involves fixing a prototype or "average" face by mapping a large number of faces, defining various landmarks on these maps, and averaging the "values" of these landmarks. This prototype face can then be used to construct

a face-space, where each point in the space identifies a distinct face. Crucially, this allows one to define the linear physical differences between faces, such that two distinct pairs of faces can physically differ by the same amount. For example, a male face can differ the same amount from a gender neural face as the gender neutral face differs from a female face.

Using this method of prototype-based transformation, one can define the average male and female face.[8] It also allows one to create "hyper-gendered" faces. Think of gender as a trajectory through this n-dimensional space, with prototypical instances of gender on each end. If one follows that trajectory beyond the gendered prototype points, one can generate hyper-male and hyper-female faces. Bestelmeyer and colleagues find repulsive aftereffects when participants adapt to male faces and subsequently view typical female faces, but no aftereffect when participants adapt to typical female faces and subsequently view hyper-female faces, even when the physical differences between male and (typical) female are the same as the differences between female and hyper-female faces. Because we only find adaptation to faces within an arbitrary region of the face space, this provides good reason to think that the aftereffects are not the result of adaptation to low-level physical differences. Instead, this suggests a specialized representational schema organized along the dimension of gender.

Is there evidence pointing to these adaptation effects being perceptual? There are very few, if any, studies that directly address this question. Theorists nearly uni-

[8]The experimenters used 20 images of white, young adult males and females to create the prototype. Of course, different prototypes would have been constructed if they had, for example, used images of black or elderly individuals.

versally assume that when participants report an aftereffect, the underlying mechanism is perceptual in nature, and rarely offer arguments for why they think it's perceptual (beyond the reference to adaptation traditionally being understood as a perceptual phenomenon). There is, however, indirect evidence that face adaptation, in general, is perceptual in nature. The argument here concerns how theorists model face adaptation, and I suspect considerations of this sort motivate claims that face adaptation is perceptual.

Models of face adaptation posit prototype representational spaces (like the one just described) to explain visual recognition and adaptation effects. The dimensions of the faces generally correspond to structural features of faces (Valentine, 1991). That is, each point in the prototype space encodes a structural description of a particular face, such that imagistic spatial relations are preserved. For the spaces constructed by experimenters, this is no accident, as the dimensions are measured from photographs (headshots taken from a particular perspective). However, if the representational structures responsible for aftereffects possess these same (or similar) dimensions, then this would give us reason to think that they are perceptual in nature.

And we have evidence that face-spaces underlie facial perception. Biologically realistic models of norm-based encoding schemes been around for a while, and are often used to explain face adaptation aftereffects (Giese and Leopold, 2005). The prevailing view is that prolonged or repeated exposure to a particular face or set of faces within a category will temporarily shift the average or prototype point. For example, adapting to a male face will pull the average (and therefore "neutral point")

towards the male end of the spectrum, such that an ambiguous face will appear female. Note, here, that the shifting of the prototype point can be understood as a Bayesian updating of the system's prior, or expected stimulus. Adaptation, then, can be understood as a temporary updating of the prior.

Mattar et al. (2016) ran an experiment investigating whether adaptation and face-space representational schemata should be understood as a single mechanism—i.e., whether face-spaces underlie the adaptation effects found at the psychophysical level. They assumed that prototypes update in a Bayesian fashion. One important variable for updating algorithms is the"stimulus history"—the window of time over which the prototype incorporates evidence. They drew on earlier work that shows that one can compare the time scale of adaptation (as measured by single cell recordings or fMRI) and the length of "stimulus history" (Hasson et al., 2008). They reasoned that if the time scale of adaptation mirrored the timescale required for prototype updating, then a single mechanism underlies both phenomenon.

To measure the time course for prototype updating, they constructed a model of prototype updating that incorporated sensory evidence over different time scales. They found that if a face space incorporates evidence over a long period (i.e., if the prototype averages stimuli over a long period), then there is very little shift after a subsequent adaption period. This is because the adapting stimulus is just one data point among many. However, if a face-space incorporates evidence over a very short period, then there are far fewer data points to average over, and the prototype shift should be pronounced. In fact, in the extreme case, the new average will be identical to value of the adapting stimulus, as the adapting stimulus is the first data point

used to calculate the face-space mean. Psychophysical studies reveal that adapting to a face moves the prototype point an intermediate amount.

Mattar and colleagues, then, had participants undergo an adaptation procedure while in the scanner. Taking measurements from the "fusiform face area" (an area within the STS), the researchers found the the time course of adaptation to faces matched the time stimulus history interval required to model the psychophysical shifts in face-space. This suggests that face-space representational schemas underlie visual adaptation to faces.

Note that this study did not look at facial categories, but only adaptation to individual faces. Thus, it could be the case that adaptation to facial categories involves post-perceptual systems. But it's unclear what would motivate this hypothesis. The study provides converging evidence that face-space representational schemas account for adaptation aftereffects, and that the mechanism is localized in areas standardly understood as visual areas. Given the similarity between the various forms of face adaptation, the most plausible explantation for race and gender aftereffects is that they are visual in nature. We don't have demonstrative proof for this hypothesis, but the current evidence strongly points us perceptually representing race and gender categories.

## 7.3 Conclusion: Closing the Representational Gap

In this chapter, I have argued for the existence of high-level visual schemata on the grounds that we find evidence of categorical perceptual structure. We do

not merely represent the structural and chromatic features of the environment. At the very least, I find sound evidence to support that we represent causal relations, animacy, and facial categories.

The idea that we visually represent abstract properties is a little hard to digest for some. It seems quite obvious that we represent colors and shape, but it remains rather mysterious how we visually represent causation, animacy, or sadness. And one might find my talk of "organizing the sensory input" to be impressionistic, at best, and hand-wavy, at worst. So allow me to illuminate one way of thinking about what these schemas do.

The traditional view says that vision is image-like. When we look out at the world, we form a representation that describes the structural features of the environment. Let's suppose that this works roughly the way a modern video camera works. The sensor continuously records properties of the light landing on it, and feeds a picture-like representation into a short-term memory buffer in which various interpretation processes can take place. Modern digital cameras include tracking technology that allow the camera to track objects through time (and in some cases single out face-like objects). Users can "lock" onto an object and the camera automatically refocuses as the object moves through the environment.

The tracking algorithms, then, affect the subsequently stored image (because they determine which object the camera will focus on), but their outputs aren't encoded into the image. However, there's no reason why they couldn't be encoded, and then passed on to the memory buffer. Imagine, then, that a camera is recording a continuously updated bitmap of one billiard ball colliding into another. The cam-

era's onboard tracking technology is specialized for, among other things, tracking the spatiotemporal dynamics that are typical of such interactions. Imagine further that these spatiotemporal dynamics are encoded and passed on in tandem with the bitmap representation to the buffer. Other systems, then, have ready access to an already interpreted "bundle" of information about the event that just took place between the two billiard balls. This proposed causal schema "organizes" the sensory stream by packaging causally-relevant spatiotemporal relations. The resulting perceptual representation is both spatially concrete and abstract, in that it is about a special class of interactions.

This analogy calls to mind Kant's original motivation for positing schemata. In his chapter on "The Schematism," Kant (1996) argues that that "empirical intuitions" (i.e., sensory perception) are fundamentally different from "pure concepts" in that the content of sensory perceptions of ordinary objects are incongruous ("heterogeneous" as Kant put it) with the content of pure concepts, which we apply to these objects. Pure concepts are highly abstract, such as UNITY, PLURALITY, SUBSTANCE, and CAUSALITY.

For less abstract geometrical concepts, Kant can envision how they might get applied to empirical objects. When we see a dinner plate, for example, we can apply the concept ROUND because "the roundness in the concept of the plate can be intuited [also] in the circle" (Kant, 1996, A 137/B176). But for pure concepts like CAUSATION or SUBSTANCE, there is no commonality between the concept and the perceptual representation.

The problem, then, is to explain how we predicate abstract concepts to em-

pirical objects when our initial sensory contact with them is devoid of properties we attribute:

> How, then, can an intuition be *subsumed* under a category, and hence how can a category be *applied* to appearances—since surely no one will say that a category (e.g., causality) can also be intuited through senses and is contained in the appearances? (A137/B176)

Kant continues, arguing that than an intermediate representation must link the sensory and the conceptual:

> Now clearly there must be something that is third, something that must be homogeneous with the category, on the one hand, and with the appearance, on the other hand, and that thus makes possible the application of the category to the appearance. This mediating presentation must be pure (i.e., without any empirical), and yet must be both *intellectual*, on the one hand, and *sensible*, on the other hand. Such a presentation is the *transcendental schema* (A177/B138).

So according to Kant, there's a representational gap between imagistic sensory representations and our pure understanding of object (i.e., our conceptual understanding). And in order to bridge that gap, there must be some sort of intermediate mental representation. Hence, empirical cognition requires both sensory-based and abstract representations.

If I understand Kant correctly, the greater the diversity of things falling under a concept, the more abstract the concept must be. And the greater the abstractness

of a concept, the greater the need for an intermediary schemata of particular types. The schema discussed in this chapter all appear to bridge a representational gap in this sense. Consider causality. We find deterministic and probabilistic causal relations; mechanical and action-at-a-distance causal relations.

The problem with Kant's line of reasoning is that we have many (many) abstract concepts. I have concepts for "mixers" (food mixers, cement mixers, paint mixers, etc.), "door knobs," "chairs," and so on. If it's a law of cognition that abstract concepts require visual schemata, then we have very large number of schemata. And this seems implausible.

But notice that Kant might be correct that there needs to be *something* that explains how sufficiently abstract concepts get applied to perceived objects. But the mechanism bridging that gap needn't always be a perceptual schemata. There may be other ways of bridging the representational gap. But for the schemata described here, it's a plausible working hypothesis that their principle role is to bridge the gap between sentience and sapience.

# Bibliography


Aguirre, G. K., Singh, R., and D'Esposito, M. (1999). Stimulus inversion and the responses of face and object-sensitive cortical areas. *Neuroreport*, 10(1):189–94.

Alais, D., Cass, J., O'Shea, R. P., and Blake, R. (2010). Visual sensitivity underlying changes in visual consciousness. *Current biology : CB*, 20(15):1362–7.

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., and De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, 23(15):1427–1431.

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *The Journal of neuroscience*, 30(8):2960–2966.

Antony, L. (2011). The Openness of Illusions. *Philosophical Issues*, 21(1):25–44.

Aristotle (1984). On the Soul. In Barnes, J., editor, *The Complete Works of Aristotle*, pages 1405–1518. Princeton University Press, Princeton, NJ.

Arrighi, R., Togoli, I., and Burr, D. C. (2014). A generalized sense of number. *Proceedings of the Royal Society B: Biological Sciences*, 281(1797):20141791–20141791.

Atlick, J. J., Li, Z., and Redlich, A. N. (1993). What Does Post-Adaptation Color Appearance Reveal About Cortical Color Representation. *Vision Research*, 33(1):123–129.

Balcetis, E. and Dale, R. (2007). Conceptual set as a top-down constraint on visual object identification. *Perception*, 36:581–596.

Bar, M. (2004). Visual objects in context. *Nature reviews. Neuroscience*, 5(8):617–29.

Barrett, H. C. and Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological review*, 113(3):628.

Barth, H., Kanwisher, N., and Spelke, E. (2003). The construction of large number representation in adults. *Cognition*, 86:201 – 221.

Bayne, T. (2009). Perception and the Reach of Phenomenal Content. *The Philosophical Quarterly*, 59(236):385–404.

Bayne, T. and Montague, M., editors (2012). *Cognitive Phenomenology*. Oxford University Press.

Beale, J. M. and Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, 57(3):217–239.

Bell, J., Gheorghiu, E., and Kingdom, F. a. a. (2009). Orientation tuning of curvature adaptation reveals both curvature-polarity-selective and non-selective mechanisms. *Journal of vision*, 9(12):3.1–11.

Benucci, A., Saleem, A. B., and Carandini, M. (2013). Adaptation maintains population homeostasis in primary visual cortex. *Nature neuroscience*, 16(6):724–9.

Bermúdez, J. (1995). Nonconceptual Content: From Perceptual Experience to Subpersonal Computational States. *Mind and Language*, 10:402–369.

Bermúdez, J. (2003). *Thinking Without Words*. Oxford University Press, New York.

Bestelmeyer, P. E. G., Jones, B. C., DeBruine, L. M., Little, a. C., Perrett, D. I., Schneider, a., Welling, L. L. M., and Conway, C. a. (2008). Sex-contingent face aftereffects depend on perceptual category rather than structural encoding. *Cognition*, 107:353–365.

Biederman, I. (1972). Perceiving Real-World Scenes. *Science*, 177(4043):77–80.

Bimler, D. and Kirkland, J. (2001). Categorical perception of facial expressions of emotion: Evidence from multidimensional scaling. *Cognition & Emotion*, 15(5):633–658.

Blakemore, S.-J., Boyer, P., Pachot-Clouard, M., Meltzoff, a., and Decety, J. (2003). The detection of contingency and animacy in the human brain. *Cerebral Cortex*, 13(8):837–844.

Block, N. (2014). Seeing-As in the Light of Vision Science. *Philosophy and Phenomenological Research*, pages 1–13.

Briscoe, R. (2015). Cognitive Penetration and the Reach of Phenomenal Content. In Zeimbekis, J. and Raftopoulos, A., editors, *The Cognitive Penetrability of Perception: New Philosophical Perspectives*, pages 174–199. Oxford University Press, New York.

Brogaard, B. (2013). Do we perceive natural kind properties? *Philosophical Studies*, 162(1):35–42.

Brovold, A. and Grush, R. (2012). Towards an (Improved) Interdisciplinary Investigation of Demonstrative Reference. In Raftopoulos, A. and Machamer, P., editors, *Perception, Realism, and the Problem of Reference*, pages 11–42. Cambridge University Press, Cambridge.

Burge, J., Peterson, M. a., and Palmer, S. E. (2005). Ordinal configural cues combine with metric disparity in depth perception. *Journal of vision*, 5(6):534–542.

Burge, T. (2010). *Origins of Objectivity*. Oxford University Press, Oxford.

Burge, T. (2011). Border crossings: Perceptual and post- perceptual object representation. *Behavioral and Brain Sciences*, 34(3):125.

Burr, D. and Ross, J. (2008). A visual sense of number. *Current biology*, 18(6):425–8.

Butler, A., Oruc, I., Fox, C. J., and Barton, J. J. S. (2008). Factors contributing to the adaptation aftereffects of facial expression. *Brain research*, 1191:116–26.

Butterfill, S. A. (2009). Seeing Causings and Hearing Gestures. *Philosophical Quarterly*, 59(236):405–428.

Calder, A. J. and Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature reviews. Neuroscience*, 6(8):641–51.

Campbell, J. and Burke, D. (2009). Evidence that identity-dependent and identity-independent neural populations are recruited in the perception of five basic emotional facial expressions. *Vision research*, 49(12):1532–40.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press, New York.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525.

Carrasco, M., Ling, S., and Read, S. (2004). Attention alters appearance. *Nature neuroscience*, 7(3):308–13.

Carruthers, P. (2006). *The Architecture of Mind*. Clarendon Press, Oxford.

Carruthers, P. (2011). *Opacity of Mind*. Oxford University Press, Oxford.

Carruthers, P. (2015). Perceiving mental states. *Consciousness and Cognition*.

Castelli, F., Happé, F., Frith, U., and Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, 12(3):314–25.

Churchland, P., Ramachandran, V., and Sejnowski, T. J. (1994). A Critique of Pure Vision. In Sejnowski, T. J., Koch, C., and Davis, J. L., editors, *Large-Scale Neuronal Theories of the Brain*, pages 23–60. Bradford, Cambridge, MA.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge University Press, Cambridge.

Churchland, P. M. (1998). Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor. *Philosophy of Science*, 55(2):167–187.

Clark, A. (2000). *A Theory of Sentience*. Clarendon Press, Oxford.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03):181–204.

Clifford, C. W. G. (2002). Perceptual adaptation: Motion parallels orientation. *Trends in Cognitive Sciences*, 6(3):136–143.

Clifford, C. W. G., Webster, M. a., Stanley, G. B., Stocker, A. a., Kohn, A., Sharpee, T. O., and Schwartz, O. (2007). Visual adaptation: neural, psychological and computational aspects. *Vision research*, 47(25):3125–31.

Craver, C. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Oxford.

Dalton, P. and Wysocki, C. J. (1996). The nature and duration of adaptation following long-term odor exposure. *Perception & psychophysics*, 58(5):781–792.

Davenport, J. L. and Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, 15(8):559–564.

de Gelder, B., Teunisse, J.-P., and Benson, P. J. (1997). Categorical perception of facial expressions: categories and their internal structure. *Cognition & Emotion*, 11(1):1–23.

Deen, B., Koldewyn, K., Kanwisher, N., and Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, 25(11):4596–4609.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, Oxford.

Descartes (1984). *The Philosophical Writings of Descartes*. Cambridge University Press, Cambridge.

Dretske, F. (1969). *Seeing and Knowing*. University of Chicago Press, Chicago.

Dretske, F. (1999). *Knowledge and the Flow of Information*. CSLI Publications, Stanford.

Dretske, F. (2000). Simple Seeing. In *Perception, Knowledge and Belief: Selected Essays*, pages 97–112. Cambridge University Press, Cambridge.

Dretske, F. (2003). Experience as Representation. *Philosophical Issues*, 13:67–82.

Dretske, F. (2015). Perception versus Conception: The Goldilocks Test. In Zeimbekis, J. and Raftopoulos, A., editors, *The Cognitive Penetrability of Perception: New Philosophical Perspectives*, volume 15, pages 163–173. Oxford University Press, New York.

Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology*, 18(18):855–856.

Eatock, R. A. (2000). Adaptation in Hair Cells. *Annual review of neuroscience*, 23:285–314.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314.

Firestone, C. and Scholl, B. J. (2014). "Top-down" effects where none should be found: the El Greco fallacy in perception research. *Psychological science*, 25(1):38–46.

Firestone, C. and Scholl, B. J. (2015). Cognition Does Not Affect Perception: Evaluating the Evidence for 'Top-Down' Effects. *Behavioral and Brain Sciences*, Online.

Fish, W. (2013). High-level properties and visual experience. *Philosophical Studies*, 162(1):43–55.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

Fodor, J. (1984). Observation Reconsidered. *Philosophy of Science*, 51(1):22–43.

Fodor, J. (2003). The Revenge of the Given. In Gunther, Y. H., editor, *Essays on Nonconceptual Content*, pages 105–116. MIT Press, Cambridge, MA.

Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press, Oxford.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press, Cambridge, MA.

Frege, G. (1948). Sense and Reference. *Philosophical Review*, 57(3):209–230.

Gao, T., McCarthy, G., and Scholl, B. J. (2010). The wolfpack effect. Perception of animacy irresistibly influences interactive behavior. *Psychological science*, 21(12):1845–53.

Gao, T., Newman, G. E., and Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59(2):154–179.

Gao, T. and Scholl, B. J. (2010). Are objects required for object-files? Roles of segmentation and spatiotemporal continuity in computing object persistence. *Visual Cognition*, 18(1):82–109.

Gao, T., Scholl, B. J., and Mccarthy, G. (2012). Dissociating the Detection of Intentionality from Animacy in the Right Posterior Superior Temporal Sulcus. *The Journal of neuroscience*, 32(41):14276–14280.

Gibson, J. (1984). *The Ecological Approach to Visual Perception*. Taylor and Francis, New York.

Gibson, J. and Radner, M. (1935). Adaptation, After-Effect and Contrast in the Perception of Tilted Lines. *Journal of Experimental Psychology*, 20(5):453–467.

Giese, M. A. and Leopold, D. A. (2005). Physiologically inspired neural model for the encoding of face spaces. *Neurocomputing*, 65-66(SPEC. ISS.):93–101.

Gilbert, C. D. and Wiesel, T. N. (1992). Receptive field dynamics in adult primary visual cortex. *Nature*, 356(6365):150–152.

Goldstone, R. L. and Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78.

Gordon, I. E., Day, R. H., and Stecher, E. J. (1990). Perceived causality occurs with stroboscopic movement of one or both stimulus elements. *Perception*, 19(1):17–20.

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23.

Grill-Spector, K. and Kanwisher, N. (2005). Visual recognition: As Soon as You know It Is There, You Know What It Is. *Psychological Science*, 16(2):152–160.

Grill-Spector, K. and Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature reviews. Neuroscience*, 15(8):536–548.

Gutteling, T. P., Petridou, N., Dumoulin, S. O., Harvey, B. M., Aarnoutse, E. J., Kenemans, J. L., and Neggers, S. F. W. (2015). Action Preparation Shapes Processing in Early Visual Cortex. *Journal of Neuroscience*, 35(16):6472–6480.

Haney, C. and Zimbardo, P. (1998). The Past and Future of U.S. Prison Policy: Twenty-Five Years After the Stanford Prison Experiment. *American Psychologist*, 53(7):709–727.

Hanson, N. (1965). *Patterns of Discovery*. Cambridge University Press, Cambridge.

Harnad, S. (2003). Categorical Perception.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of neuroscience*, 28(10):2539–50.

Haynes, J.-D. and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5):686–691.

Hegdé, J. and Kersten, D. (2010). A link between visual disambiguation and visual memory. *The Journal of neuroscience*, 30(45):15124–15133.

Heider, F. and Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, 57(2):243–259.

Hilgetag, C. C., Burns, G. a., O'Neill, M. a., Scannell, J. W., and Young, M. P. (2000). Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 355(1393):91–110.

Hohwy, J. (2014). *The Predictive Brain*. Oxford University Press, Oxford.

Howe, C. Q. and Purves, D. (2005). The Muller-Lyer illusion explained by the statistics of image source relationships. *Proceedings of the National Academy of Sciences*, 102(4).

Hsieh, P.-J., Vul, E., and Kanwisher, N. (2010). Recognition alters the spatial pattern of FMRI activation in early retinotopic cortex. *Journal of neurophysiology*, 103(3):1501–1507.

Javadi, A. H. and Wee, N. (2012). Cross-Category Adaptation: Objects Produce Gender Adaptation in the Perception of Faces. *PLoS ONE*, 7(9):3–10.

Jiang, J., Summerfield, C., and Egner, T. (2013). Attention sharpens the distinction between expected and unexpected percepts in the visual brain. *The Journal of neuroscience*, 33(47):18438–47.

Kahneman, D., Treisman, A., and Gibbs, B. J. (1992). The reviewing of object-files: Object specific integration of information. *Cognitive Psychology*, 24:174–219.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685.

Kant, I. (1996). *Critique of Pure Reason*. Hackett, Indianapolis, IN.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America*, 107:11163–11170.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 17(11):4302–11.

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of neurophysiology*, 97(5):3155–64.

Kok, P., Brouwer, G. J., van Gerven, M. a. J., and de Lange, F. P. (2013). Prior Expectations Bias Sensory Representations in Visual Cortex. *Journal of Neuroscience*, 33(41):16275–16284.

Kok, P. and De Lange, F. P. (2014). Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. *Current Biology*, 24(13):1531–1535.

Kok, P., Failing, M. F., and de Lange, F. P. (2014). Prior Expectations Evoke Stimulus Templates in the Primary Visual Cortex. *Journal of Cognitive Neuroscience*, 26(7):1546–1554.

Kok, P., Jehee, J. F. M., and de Lange, F. P. (2012a). Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron*, 75(2):265–270.

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., and De Lange, F. P. (2012b). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex*, 22(9):2197–2206.

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

Krekelberg, B., Boynton, G. M., and van Wezel, R. J. a. (2006). Adaptation: from single cells to BOLD signals. *Trends in neurosciences*, 29(5):250–6.

Lee, S. H., Kravitz, D. J., and Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *NeuroImage*, 59(4):4064–4073.

Leopold, D. A. and Logothetis, N. K. (1999). Multistable phenomena : changing views in perception. *Trends in cognitive sciences*, 3(7):254–264.

Leopold, D. a., Rhodes, G., Müller, K.-M., and Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proceedings. Biological sciences / The Royal Society*, 272(1566):897–904.

Li, W., Piëch, V., and Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience*, 7(6):651–657.

Little, A. C., DeBruine, L. M., and Jones, B. C. (2005). Sex-contingent face aftereffects suggest distinct neural populations code male and female faces. *Proceedings. Biological sciences / The Royal Society*, 272(1578):2283–7.

Logue, H. (2013). Visual experience of natural kind properties: is there any fact of the matter? *Philosophical Studies*, 162(1):1–12.

Long, G. M. and Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: alternating views of reversible figures. *Psychological bulletin*, 130(5):748–768.

Macpherson, F. (2012). Cognitive Penetration of Colour Experience : Rethinking the Issue in Light of an Indirect Mechanism. *Philosophy and Phenomenological Research*, 84(1):24–62.

Mandelbaum, E. (2016). Seeing and Conceptualizing: Modularity and the Shallow Contents of Perception. *Philosophy and Phenomenological Research*.

Markov, N. T., Ercsey-Ravasz, M., Van Essen, D. C., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013). Cortical High-Density Counterstream Architectures. *Science*, 342(6158):1238406–1238406.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, Cambridge, MA.

Mattar, M. G., Kahn, D. A., Thompson-Schill, S. L., and Aguirre, G. K. (2016). Varying Timescales of Stimulus Integration Unite Neural Adaptation and Prototype Formation. *Current Biology*, 26(13):1669–1676.

McGinn, C. (1982). *The Character of Mind*. Oxford University Press, Oxford.

Mechelli, A., Price, C. J., Friston, K. J., and Ishai, A. (2004). Where bottom-up meets top-down: Neuronal interactions during perception and imagery. *Cerebral Cortex*, 14(11):1256–1265.

Michotte, A. (1963). *The Perception of Causality*. Basic Books, New York.

Moulton, S. T., Kosslyn, S. M., and B, P. T. R. S. (2009). Imagining predictions: mental imagery as mental emulation. *Philosophical Transactions of the Royal Society*, 364:1273–1280.

Murray, E. a., Gaffan, D., and Mishkin, M. (1993). Neural substrates of visual stimulus-stimulus association in rhesus monkeys. *The Journal of neuroscience*, 13(10):4549–4561.

Musall, S., von der Behrens, W., Mayrhofer, J. M., Weber, B., Helmchen, F., and Haiss, F. (2014). Tactile frequency discrimination is enhanced by circumventing neocortical adaptation. *Nature Neuroscience*, 17(11).

Neander, K. (1995). Misrepresenting and Malfunction. *Philosophical Studies*, 79(2):109–141.

Nelson, D. and Marler, P. (1989). Categorical Perception of a Natural Stimulus Continuum: Birdsong. *Science*, 244(4907):976–978.

Neri, P. (2014). Semantic control of feature extraction from natural scenes. *The Journal of neuroscience*, 34(6):2374–88.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

Ogilvie, R. and Carruthers, P. (2016a). Firestone & Scholl conflate two distinct issues. *Behavioral and Brain Sciences*.

Ogilvie, R. and Carruthers, P. (2016b). Opening up Vision: The case against encapsulation. *Review of Philosophy and Psychology*, 7(4):721–742.

Orlandi, N. (2014). *The Innocent Eye: Why Vision is Not a Cognitive Process*. Oxford University Press, New York.

Osvath, M. and Karvonen, E. (2012). Spontaneous innovation for future deception in a male Chimpanzee. *PLoS ONE*, 7(5).

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3(5):519–526.

Palmer, S. E. (1999). *Vision science: From Photons to Phenomenology*. MIT Press, Cambridge, MA.

Papenmeier, F., Meyerhoff, H. S., Jahn, G., and Huff, M. (2014). Tracking by location and features: object correspondence across spatiotemporal discontinuities during multiple object tracking. *Journal of experimental psychology. Human perception and performance*, 40(1):159–71.

Patel, A. (2008). *Music, Language, and the Brain*. Oxford University Press, Oxford.

Payne, B. K. (2001). Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of personality and social psychology*, 81(2):181–192.

Peacocke and Christopher (1992). *A Study of Concepts*. MIT Press, Cambridge, MA.

Peacocke, C. (1998). Nonconceptual Content Defended. *International Phenomenological Society*, 58(2):381–368.

Pérez-González, D. and Malmierca, M. S. (2014). Adaptation in the auditory system: an overview. *Frontiers in integrative neuroscience*, 8(February):19.

Piazza, M., Izard, V., Pinel, P., Bihan, D. L., and Dehaene, S. (2004). Tuning Curves for Approximate Numerosity in the Human Intraparietal Sulcus. *Neuron*, 44:547–555.

Piazza, M., Pinel, P., Le Bihan, D., and Dehaene, S. (2007). A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal Cortex. *Neuron*, 53(2):293–305.

Pinker, S. (1997). *How the Mind Works*. Norton, New York.

Pinto, Y., Gaal, S. V., Lange, F. P. D., Lamme, V. a. F., and Seth, A. K. (2015). Expectations accelerate entry of visual stimuli into awareness. *Journal of Vision*, 15(8):1–15.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research*, 41:3145–3161.

Pond, S., Kloth, N., McKone, E., Jeffery, L., Irons, J., and Rhodes, G. (2013). After-effects Support Opponent Coding of Face Gender. *Journal of Vision*, 13(14):16.

Potter, M. (1975). Meaning in Visual Search. *Science*, 187(4180):965–966.

Prinz, J. (2006). Is the Mind Really Modular. In Stainton, R. J., editor, *Contemporary Debates in Cognitive Science*, pages 22–36. Blackwell, Malden, MA.

Purves, D. and Lotto, R. B. (2011). *Why We See What We Do Redux: A Wholly Emprical Theory of Vision*. Sinauer, Sunderland, MA.

Pylyshyn, Z. (1999). Is Vision Continuous with Cognition? *Behavioral and Brain Sciences*, 22:341–423.

Pylyshyn, Z. (2003). Return of the mental image: Are there really pictures in the brain? *Trends in Cognitive Sciences*, 7(3):113–118.

Pylyshyn, Z. (2007). *Things and Places: How the Mind Connects with the World*. Bradford, Cambridge, MA.

Pylyshyn, Z. W. (1980). Computation and cognition: issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3:111–169.

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, 4(5):197–207.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1-2):127–158.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. *Philosophical Review*, 60(1):20–43.

Ramachandran, V. and Anstis, S. (1986). The Perception of Apparent Motion. *Scientific American*, 254(6):102–109.

Ramalingam, N., Mcmanus, J. N. J., Li, W., and Gilbert, C. D. (2013). Top-Down Modulation of Lateral Interactions in Visual Cortex. *The Journal of neuroscience*, 33(5):1773–1789.

Ramus, F., Ramus, F., Hauser, M. D., Miller, C., Morris, D., and Mehler, J. (2008). Language Discrimination by Human Newborns and by Cotton-Top Tamarin Monkeys. *Science*, 349(2000):349–351.

Reddy, L., Tsuchiya, N., and Serre, T. (2010). Reading the mind's eye: Decoding category information during mental imagery. *NeuroImage*, 50(2):818–825.

Rémy, F., Saint-Aubert, L., Bacon-Macé, N., Vayssière, N., Barbeau, E., and Fabre-Thorpe, M. (2013). Object recognition in congruent and incongruent natural scenes: A life-span study. *Vision Research*, 91:36–44.

Rey, G. (1998). A Narrow Representationalist Account of Qualitative Experience. *Philosophical Perspectives*, 12:435–457.

Rey, G. (2014). Analytic, A Priori, False - And Maybe Non-Conceptual. *European Journal of Analytic Philosophy*, 10(2).

Rips, L. J. (2011). Causation From Perception. *Perspectives on Psychological Science*, 6(1):77–97.

Rolfs, M., Dambacher, M., and Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current biology*, 23:250–4.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Russell, B. (1910). Knowledge by Acquaintance and Knowledge by Description. *Proceedings of the Aristotelian Society*, 11:108–128.

Russell, B. (1912). *The Problems of Philosophy*. Dover, Mineola, NY.

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., and Kastner, S. (2012). The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands. *Science*, 337(August):753–756.

Sanchez-Vives, M. V., Nowak, L. G., and McCormick, D. a. (2000). Cellular mechanisms of long-lasting adaptation in visual cortical neurons in vitro. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20(11):4286–4299.

Sato, W., Kochiyama, T., Yoshikawa, S., Naito, E., and Matsumura, M. (2004). Enhanced neural activity in response to dynamic facial expressions of emotion: An fMRI study. *Cognitive Brain Research*, 20(1):81–91.

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., and Turk-Browne, N. B. (2014). The Necessity of the Medial Temporal Lobe for Statistical Learning. *Journal of Cognitive Neuroscience*, 26(8):1736–1747.

Schlottmann, A., Ray, E. D., Mitchell, A., and Demetriou, N. (2006). Perceived physical and social causality in animated motions: spontaneous reports and ratings. *Acta psychologica*, 123(1-2):112–43.

Scholl, B. J. (2001). Objects and Attention: The State of the Art. *Cognition*, 80:1–46.

Scholl, B. J. (2007). Object persistence in philosophy and psychology. *Mind and Language*, 22(5):563–591.

Scholl, B. J. and Gao, T. (2013). Perceiving Animacy and Intentionality: Visual Processing or Higher-Level Judgment. In Rutherford, M. D. and Kuhlmeier, V. A., editors, *Social Perception*, pages 197–229. MIT Press, Cambridge, MA.

Scholl, B. J. and Nakayama, K. (2002). Causal Capture: Contextual Effects on the Perception of Collision Events. *Psychological Science*, 13(6):493–498.

Scholl, B. J. and Nakayama, K. (2004). Illusory causal crescents: Misperceived spatial relations due to perceived causality. *Perception*, 33(4):455–469.

Scholl, B. J. and Tremoulet, P. D. (2000). Perceptual causality and animacy.

Schultz, J., Imamizu, H., Kawato, M., and Frith, C. D. (2004). Activation of the human superior temporal gyrus during observation of goal attribution by intentional objects. *Journal of Cognitive Neuroscience*, 16(10):1695–1705.

Schweinberger, S. R., Casper, C., Hauthal, N., Kaufmann, J. M., Kawahara, H., Kloth, N., Robertson, D. M. C., Simpson, A. P., and Zäske, R. (2008). Auditory Adaptation in Voice Perception. *Current Biology*, 18(9):684–688.

Schwiedrzik, C. M., Ruff, C. C., Lazar, A., Leitner, F. C., Singer, W., and Melloni, L. (2014). Untangling perceptual memory: hysteresis and adaptation map into separate cortical networks. *Cerebral Cortex*, 24(5):1152–64.

Shea, N. (2015). Distinguishing Top-Down From Bottom-Up Effects. In Stokes, D. and Matthen, M., editors, *Perception and Its Modalities*, pages 73–91. Oxford University Press, Oxford.

Shepard, R. N. and Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, 171(3972):701–703.

Siegel, S. (2010). *The Contents of Visual Experience*. Oxford University Press, Oxford.

Slotnick, S. D., Thompson, W. L., and Kosslyn, S. M. (2012). Visual memory and visual mental imagery recruit common control and sensory regions of the brain. *Cognitive Neuroscience*, 3(1):14–20.

Solomon, S. G. and Kohn, A. (2014). Moving sensory adaptation beyond suppressive effects in single neurons. *Current Biology*, 24(20):R1012–R1022.

Spelke, E. S. (1988). Where perceiving ends and thinking begins: The apprehension of objects in infancy. In Yonas, A., editor, *Perceptual development in infancy: The Minnesota Symposia on Child Psychology*, pages 197–234. Lawrence Erlbaum Associates.

Sperber, D. (2002). In Defense of Massive Modularity. In Dupoux, I., editor, *Language, Brain, and Cognitive Development*, pages 47–57. MIT Press, Cambridge, MA.

St. John-Saaltink, E., Utzerath, C., Kok, P., and Lau, H. C. (2015). Expectation Suppression in Early Visual Cortex Depends on Task Set. *PloS ONE*, 10(6):1–14.

Stone, J. V. (2011). Footprints sticking out of the sand. Part 2: Children's Bayesian priors for shape and lighting direction. *Perception*, 40(2):175–190.

Summerfield, C. and de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(November).

Summerfield, C. and Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, 13(9):403–409.

Susilo, T., McKone, E., and Edwards, M. (2010). Solving the upside-down puzzle : Why do upright and inverted face aftereffects look alike ? *Journal of Vision*, 10(13):1–16.

Sweeny, T. D., Haroz, S., and Whitney, D. (2012). Reference repulsion in the categorical perception of biological motion. *Vision Research*, 64:26–34.

Teufel, C., Subramaniam, N., and Fletcher, P. C. (2013). The role of priors in Bayesian models of perception. *Frontiers in computational neuroscience*, 7(April):25.

Thomas, R., Nardini, M., and Mareschal, D. (2010). Interactions between "light-from-above" and convexity priors in visual development. *Journal of Vision*, 10(8):6.1–7.

Thorpe, S., Fize, D., Marlot, C., and Marlo, C. (1996). Speed of processing in the human visual system. *Letters to Nature*, 381(6582):520–2.

Tiddeman, B., Burt, M., and Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, 21(5):42–50.

Tong, F. (2013). Imagery and visual working memory: One and the same? *Trends in Cognitive Sciences*, 17(10):489–490.

Treisman, A. M. and Gelade, G. (1980). A Feature-integration Theory of Attention. *Cognitive Psychology*, 12:97–136.

Tremoulet, P. D. and Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, 29:943–951.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press, Cambridge, MA.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 43(May 2012):161–204.

Vetter, P. and Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, 27(1):62–75.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. a., Singh, M., and von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figureground organization. *Psychological Bulletin*, 138(6):1172–1217.

Webster, M. A. (2011). Adaptation and visual coding. *Journal of Vision*, 11(5):1–23.

Webster, M. A. (2015). Visual adaptation. *Annual Review of Vision Science*, 1(1):547–569.

Webster, M. A., Kaping, D., Mizokami, Y., and Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, 428:357–360.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604.

White, P. a. (2012). Visual impressions of causality: Effects of manipulating the direction of the target object's motion in a collision event. *Visual Cognition*, 20(2):121–142.

White, P. a. and Milne, A. (1997). Phenomenal Causality: Impressions of Pulling in the Visual Perception of Objects in Motion. *The American Journal of Psychology*, 110(4):573–602.

Winston, J. S. (2004). fMRI-Adaptation Reveals Dissociable Neural Representations of Identity and Expression in Face Perception. *Journal of Neurophysiology*, 92(3):1830–1839.

Wyttenbach, R. A., May, M. L., and Hoy, R. R. (1996). Categorical Perception of Sound Frequency by Crickets. *Science*, 273:1542–1544.

Young, L. R., Sienko, K. H., Lyne, L. E., Hecht, H., and Natapoff, A. (2003). Adaptation of the vestibulo-ocular reflex, subjective tilt, and motion sickness to head movements during short-radius centrifugation. *Journal of vestibular research*, 13(2-3):65–77.

Yuille, A. and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308.