Philosophical Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cphp20

Embedded seeing-as: Multi-stable visual perception without interpretation

Nicoletta Orlandi

Available online: 25 Jul 2011

To cite this article: Nicoletta Orlandi (2011): Embedded seeing-as: Multi-stable visual perception without interpretation, Philosophical Psychology, DOI:10.1080/09515089.2011.579425

To link to this article: http://dx.doi.org/10.1080/09515089.2011.579425

Philosophical Psychology
2011, 1–19, iFirst

Embedded seeing-as: Multi-stable visual perception without interpretation

Nicoletta Orlandi

Standard models of visual perception hold that vision is an inferential or interpretative process. Such models are said to be superior to competing, non-inferential views in explanatory power. In particular, they are said to be capable of explaining a number of otherwise mysterious, visual phenomena such as multi-stable perception. Multi-stable perception paradigmatically occurs in the presence of ambiguous figures, single images that can give rise to two or more distinct percepts. Different interpretations are said to produce the different percepts. In this paper, I argue that a non-inferential account of visual perception is just as capable of explaining multi-stable perception. I propose an embedded understanding of vision, and show how the embedded account can, after properly qualifying them, use the explanatory resources of the inferential view to explain just what such a view explains.

Keywords: Embedded; Inference; Interpretation; Multi-stability; Vision

Standard accounts of visual perception conceive of vision as an inferential process (Churchland, 1989; Fodor, 1988; Fodor & Pylyshyn, 1981; Gregory, 1970; Marr, 1982; Palmer, 1999; Rock, 1983), in particular as a process that moves from retinal images to representations of objects and scenes in the environment in representational stages, and by following a number of rules. I call the idea that vision involves inferences the "Standard View" (SV). The Standard View is said to be superior to competing, non-inferential accounts of vision, in explanatory power: it is said to be capable of explaining a number of otherwise mysterious visual phenomena. One such phenomenon, multi-stable perception, occurs paradigmatically in the presence of ambiguous figures, single images that can give rise to two or more percepts. Proponents of SV hold that the different percepts are the product of different interpretations.

Nicoletta Orlandi is Assistant Professor of Philosophy and Cognitive Science at Rice University.
Correspondence to: Nicoletta Orlandi, Department of Philosophy, Rice University, 6100 Main Street MS 14, Houston, Texas 77005, USA. Email: nico@rice.edu

ISSN 0951-5089 (print)/ISSN 1465-394X (online)/11/-0001-19 © 2011 Taylor & Francis
DOI: 10.1080/09515089.2011.579425

In this paper, I argue that a non-inferential and direct account of vision is just as capable of explaining multi-stable perception. I propose an embedded understanding of visual processing, according to which the visual system can rely on environmental regularities to produce representations of objects and scenes without performing inferences. I call the embedded account, the "Embedded View" (EV). I argue that EV can, after properly qualifying them, use the explanatory resources of SV and explain just what SV explains. I further suggest that we have theoretical reasons for preferring EV to SV, reasons that have to do with considerations of parsimony.

Like proponents of SV, I talk interchangeably of "inferences," "interpretations," and "constructions," and I will sometimes refer to the representational products of visual processing as "percepts." Further, in line with SV, I assume that the visual system is composed of the eye, with its retina and optic nerve, and of the visual cortex (Palmer, 1999, p. 56). I will also talk of "representations" as states of a system that carry information about some environmental entity, can misrepresent—sometimes occurring in the absence of what they carry information about—and serve in the performance of some cognitive task.[1]

In section 1, I introduce the Standard View and spell out its perceived advantages. In section 2, I present the Embedded View and clarify in detail how it differs from SV. In section 3, I show how the EV, just like SV, can account for multi-stable perception, and, in section 4, I sketch a preliminary picture of how multi-stability is explained in an embedded framework. I conclude that the explanatory advantage of SV is only apparent.

1. The Standard View

Most contemporary vision scientists hold that vision is, with proper qualifications, an intelligent process (Churchland, 1989; Fodor, 1988; Fodor & Pylyshyn, 1981; Gregory, 1970; Marr, 1982; Palmer, 1999; Rock, 1983). It is intelligent because it involves an interpretation of the stimulation present at the retina in terms of the world of objects. Palmer writes:

> The objects that we so effortlessly perceive are not the direct cause of our perceptions. Rather, perceptions are caused by the two-dimensional patterns of light that stimulate our eyes. . . . To provide us with information about the three-dimensional environment, vision must therefore be an interpretative process that somehow transforms complex, moving, two-dimensional patterns of light at the back of the eyes into stable perceptions of three-dimensional objects in three-dimensional space. We must therefore conclude that the objects we perceive are actually interpretations based on the structure of images rather than direct registrations of physical reality. (1999, p. 9)

Like Sherlock Holmes, who infers the identity of a criminal from scattered evidence, the visual system infers representations of objects and scenes from patterns of light. More specifically, visual perception proceeds in stages where low-level states of the visual system, obtained directly from retinal stimulation, represent features such as intensity values and light discontinuities; high-level states represent more

complex entities like edges and objects, and they are produced as a result of an interpretation of the information contained in the earlier states.

The number of vision scientists that accept this kind of view is substantial enough to warrant labeling this position, as I will do in the present paper, the ''Standard View'' of visual perception (SV). Proponents of SV include cognitive psychologists, philosophers and neuroscientists. Gregory, for example, writes:

> Perception involves a kind of inference from sensory data to object-reality. Further, behavior is not controlled directly by the data, but by the solutions to the perceptual inferences from the data. This is clear from common experience: if I put a book on a table I do not prod the table first to check that it is solid. I act according to the *inferred* physical object-table—not according to the brown patch in my eye. So perception involves a kind of problem-solving—a kind of intelligence. Helmholtz spoke of perception in terms of ''unconscious inferences.'' (1970, p. 30)

Along similar lines, Fodor and Pylyshyn write: ''the current Establishment theory (sometimes referred to as the information processing view) is that perception depends, in several respects presently to be discussed, upon *inferences*'' (1981, p. 140).

And neuroscientist Friston (together with his collaborators) writes:

> There is growing support of the idea that the brain is an inference machine, or hypothesis tester, which approaches sensory data using principles similar to those that govern the interrogation of scientific data. In this view, perception is a type of unconscious inference. (Hohwy, Roepstorff, & Friston, 2008, p. 2).

Given the ease and automaticity of perception, we may view the idea that the visual system performs inferences as implausible. At first impression, the inferential account seems to over-intellectualize what is involved in perceiving objects. But defenders of SV have the resources to dispel these doubts. For, first, the idea that vision is inferential is qualified by noticing the differences between inferences commonly performed by people and those performed by visual systems. The former are typically conscious, deliberate, slow, and verbal, while the latter are unconscious, effortless, rapid and non-verbal (Palmer, 1999, p. 80).

Moreover, many contemporary proponents of SV are adherents to the ''Computational Theory of Mind,'' the view that mental processes, including perceptual processes, are computational (Fodor & Pylyshyn, 1981; Marr, 1982; Palmer, 1999; Rock, 1983). This allows them to make use of the computer analogy. Perceptual systems perform inferences in the same way in which computers perform inferences: they manipulate symbols in virtue of rules (or algorithms) in fast and automatic ways. Palmer says: ''the computer analogy is quite compatible with the inferential analogy of construction because making inferences is, in effect, what computers do when they execute programs'' (1999, p. 71). Proponents of the standard view are then capable of responding to intuitive reservations about their position by pointing out that inferences can be effortless and automatic just like computations are effortless and automatic.

They further argue that their view is superior in explanatory power to competing scientific theories of perception. In particular, SV is believed to be superior to *direct* accounts of perception. Direct accounts hold that perception is not mediated by the manipulation of representations that precede the representation of objects and scenes (Gibson, 1979). Visual phenomena that are supposedly unexplained by the direct account include both the general stability of visual perception, and its instability in the presence of ambiguous figures. The latter phenomenon is of particular interest to us, but it might be useful to start by explaining how SV accounts for the stability of our visual world in the face of underdetermined visual stimulation (Fodor & Pylyshyn, 1981; Gregory, 1970, pp. 25 & 142; Palmer, 1999, p. 55). This will help us understand the solution to the multi-stability problem as well.

The idea that visual systems perform interpretations is often introduced by the consideration that there is a one-to-many correspondence between retinal images and their causes. The visual stimulus that hits the retina in the form of a pattern of light is said to be ambiguous, in the sense of being underdetermined, or compatible with a number of different distal objects and scenes. But we tend to perceive a stable world. How do we explain this fact? It is suggested that the visual system processes the stimulus by interpreting the ambiguous available data. Palmer, for example, says:

> For every 2-D image on the back of our eyes, there are infinitely many distinct 3-D environments that could have given rise to it. . . . Vision is thus a heuristic process in which inferences are made about the most likely environmental condition that could have produced a given image. (1999, p. 23)

The stimulus for vision is underdetermined, being compatible with a number of distal scenes, but the product of visual processing is not: we tend to see pretty much the same objects when we look at a given scene for a continuous period of time (or at different times). The retinal stimulus by itself does not guarantee such stability. What does? Proponents of SV suggest that the system that processes the stimulus must be responsible for it. The visual system produces stable representations of the world by bringing to bear information about it that helps to solve the underdetermination problem. Proponents of SV are then inclined to suppose that the visual system performs inferences. More specifically, they are inclined to suppose that the system solves the problem it faces by following some encoded rules that make use of assumptions concerning the make-up of objects in the world (Marr, 1982; Marr & Hildreth, 1980; Rock, 1983; Spelke, 1990; Ulman, 1979). Palmer says:

> How does [the visual system] solve this seemingly insoluble problem? Different theorists have taken different approaches . . . but the dominant one is to assume that 3-D perception results from the visual system making lots of highly plausible assumptions about the nature of the environment and the conditions under which it is viewed. These assumptions constrain the inverse problem enough to make it solvable most of the time. (1999, p. 23)

The visual system produces percepts by using some stored knowledge of the world. For example, sudden discontinuities in light intensity at the retina—what Marr called "zero-crossings"—are typically, but not invariably, caused by the presence of edges.

The visual system assumes that this is the case and produces a representation of edges from a representation of the discontinuities. Marr and Hildreth call this principle the ''spatial coincidence assumption.'' They say:

> If a zero-crossing segment is present in a set of independent channels over a contiguous range of sizes and the segment has the same position and orientation in each channel, then the set of such zero-crossing segments may be taken to indicate the presence of an intensity change in the image that is due to single physical phenomenon (a change in reflectance, illumination, depth or surface orientation).... It follows... that provided the two channels are reasonably separated in the frequency domain, and their zero-crossings agree, the combined zero-crossings can be taken to indicate the presence of an edge in the image. (Marr & Hildreth, 1980, p. 202)

At later stages of visual processing, the visual system is also said to move from the representation of surfaces in motion to the representation of objects by assuming that the causes of retinal stimuli are rigid (Ulman, 1979). The visual system knows that, roughly, surfaces bundle together to form rigid objects and, by assuming this fact, it produces representations of objects out of the representation of the surfaces.

More recent incarnations of this view bring the resources of Bayesian theory to bear on the solution to the underdetermination problem (van Ee, 2003; Hohwy et al., 2008). The idea is that the visual system forms hypotheses that predict what the sensory input should be, if it were caused by certain environmental entities. The hypotheses are then checked against the evidence, and the one that generates the best predictions determines the percept. Hohwy et al. say:

> For example, if the hypothesis is that visual input is caused by a box, then it is possible to predict, on the basis of that hypothesis, what the input is going to be as one moves around it. If the prediction turns out to be right, and if the presence of a box is otherwise probable, then the probability for the hypothesis that it is a box goes up. If there are no better hypotheses in play, then this hypothesis wins and the perceptual inference will be that the environmental cause is indeed a box. (2008, pp. 2–3)

In so far as views like the one just quoted conceive of the visual system as using worldly information to form hypotheses concerning represented retinal stimuli, they, like their non-Bayesian predecessors, tend to understand perceptual systems as possessing some stored knowledge of the world to reduce the possible interpretations of the retinal data. In other words, perceptual systems use knowledge to reduce (to pretty much just one) the representations that are produced given the stimulus. This explains the fact that, despite the compatibility of the stimulus with many distal causes, we experience a fairly unchanging world. A direct account of perception that attempts to do without positing visual inferences has trouble explaining this very fact.

SV can adopt a similar explanatory strategy to account for multi-stable perception where talk of interpretation seems very plausible. Palmer says:

> Ambiguous figures demonstrate the constructive nature of perception because they show that perceivers interpret visual stimulation and that more than one interpretation is sometimes possible. If perception were completely determined

> by the light stimulating the eye, there would be no ambiguous figures because each pattern of stimulation would map onto a unique percept. This position is obviously incorrect. Something more complex and creative is occurring in vision, going beyond the information strictly given in the light that stimulates our eyes. (1999, p. 10).

It is somewhat curious that appeal to inferences serves to explain *both* why vision is stable in the face of underdetermined retinal stimuli, and why it is *not* stable in the presence of ambiguous figures. One might think that we would need some additional explanation for why stability is violated in certain cases. We will return to this point in section 3; at present, we should agree that we have a case in favor of SV. SV is supported by the traditional Computational Theory of Mind, and it is able to explain a number of otherwise mysterious visual phenomena. In the next section, I propose an alternative way of understanding visual processing. I then show that the alternative is as explanatory powerful as SV. Elsewhere (Orlandi, unpublished manuscript), I have argued that a situated and direct account of visual processing has no trouble explaining misperception and visual illusion. Here, I take on the task of showing that it can also account for multi-stable perception.

2. The Embedded View

The key to provide an alternative to the Standard View is to explain how vision can carry out its functions without performing inferences. The first step in understanding how this is possible is to think of the visual system as *relying* on certain environmental regularities. The contrast here is between relying on the regularities versus representing them in the form of assumptions about the physical world.[2] In computational models of vision, the visual system uses some prior knowledge of the world in order to build a representation of objects out of retinal images. For example, the system assumes that objects in the world are rigid. Such assumption has to be represented or encoded somewhere within the system because it has to serve in visual computations. Similarly, in Bayesian models, the visual system forms hypotheses about the likely cause of a represented environmental stimulus by using some encoded assumption about what caused the stimulus.

Alternatively, we can think of the visual system as relying on regularities occurring in the world without encoding them. Objects in the world are typically rigid and the visual system can rely on this fact to produce a representation of objects directly form the retinal stimulation. Because the causes of retinal stimuli are typically rigid, we end up seeing the world that way. Likewise, discontinuities in light intensity at the retina can generally be ascribed to the presence of edges. The light that bounces off edges typically varies in intensity. Now, we can either think of the visual system as antecedently knowing this fact, or we can think of it as relying on this fact. The latter is the proposal I favor. The visual system doesn't first represent light discontinuities and then infer the presence of edges given its antecedent knowledge, nor does it hypothesize that an edge is present by using such knowledge: it rather relies on the fact that discontinuities are caused by edges to represent their presence.

Systems that rely on environmental regularities without knowing anything about them are ubiquitous. Fire alarms seem to assume that smoke is typically caused by fire. This is why, when they detect smoke, they signal the presence of fire. But it is implausible to describe fire alarms as actually knowing anything about such regularity. Fire alarms have been built to signal the presence of fire by relying on the fact that smoke is typically caused by it: they don't have to represent such regularity in order to perform their function. Similarly, they detect the presence of smoke by relying on additional regularities. Optical fire alarms, for example, work roughly in the following way: they emit a beam of light and have a built-in photoelectric sensor in the proximity of the beam inside the so called "optical chamber." In the absence of smoke, the light in the beam travels in a straight line. When smoke enters the optical chamber and interferes with the path of light, some light is scattered by smoke particles, directing it at the sensor which, in turn, triggers the alarm. In this case, the alarm relies on the fact that smoke particles deflect the direction of light in order to perform its operations. The device doesn't need to possess this fairly sophisticated piece of knowledge. It simply relies on the physical facts to work as it does.

Notice that in this, and many other cases, the regularities can enter into an explanation of the detector's behavior without being part of the detector's knowledge. The facts that smoke is typically caused by fire, and that smoke particles deflect light particles, explain why the device does what it does without being represented by the device.[3]

If we think of vision on the model of fire alarms, then we can stop thinking of the visual system as performing inferences: the system does not need to represent the assumptions on which it operates, it simply relies on the environment to perform its function. Correspondingly, the system doesn't need to move from low-level to object-level visual representations. It doesn't infer the one from the other. Just like states of a fire alarm, early visual states (such as retinal images) are sensitive to environmental properties, but they are not ipso facto representations. The properties of the environment that they are sensitive to serve in visual operations, but these operations do not resemble those of Sherlock Holmes, Bayesian hypothesis-testers, or digital computers: they are not manipulations of symbols in virtue of encoded rules. Derivatively, early visual states don't need to be understood as symbols or representations. For this reason, calling whatever is on the retina an "image" is misleading.[4]

Proponents of SV would, in fact, have to agree with this: in their view, early visual states are representations not just because they are sensitive to environmental features, but also because they serve in inferential operations. According to them, not any state that is sensitive to the environment is a representation: states of a smoke detector, for example, are not representations. Only those states that serve some cognitive function, e.g., computing or inferring, achieve representational status (Segal, 1989). The states produced as a result of visual processing, for example, are plausibly representations because, on top of containing information about the world, they also ground our beliefs and judgments. Early visual states, by contrast, are not thought to ground beliefs and judgments. We do not learn about light discontinuities

from our retinal states. Additionally, if the embedded view is right, such states do not serve in inferences or other similar cognitive operations. But then the need to call these states representations slowly disappears.

There is, then, a clear sense in which the embedded visual system produces representations of the world directly: the result of its processing is a schema of the environment, but this result is achieved without the mediation of further representational stages. Accordingly, the system can be *described* as performing inferences, but what it does is rather rely on certain environmental regularities to directly produce representations of edges, rigid objects and scenes out of retinal inputs.

Now, before we proceed to discuss the explanatory power of EV, let me make sure it is clear why EV is different from, and a competitor of, SV. We may start by noticing that EV is bound to ascribe less internal complexity to the visual system: in particular, it is bound to ascribe less knowledge of the world to such a system. We may think that this is in fact a good theoretical reason to prefer EV to SV. In standard models of vision, the perceptual system is said to possess information about the world that enables the solution of the underdetermination problem. Questions arise concerning the origin of such information: it is common to think that this information is innately encoded in the system because it is hard to see how the system could have acquired it (Fodor, 1983). Appeal to evolution is ineffective: evolutionary theory can explain how species acquire behavioral and physical traits such as lungs and brains, but it cannot account for the acquisition of knowledge of environmental facts. A theory of vision that posited less innate material would be preferable. EV is such a theory: according to EV, the visual system does not encode or represent the assumptions needed to interpret retinal images. It simply relies on environmental regularities to produce representations. And, quite apart from the issue of innate knowledge, EV is a more ontologically parsimonious theory than SV. Where SV posits representational states that mediate the creation of visual representations, EV doesn't. The process that issues visual representations need not involve other representations. For these reasons, if we showed that EV is as explanatory powerful as SV, we would be justified in preferring it.

Notice further that the difference between SV and EV is not merely verbal. The main difference, as I have stated it, is this: according to SV, the visual system encodes or represents the assumptions about the physical world that it needs to use in order to interpret retinal images; according to EV, it doesn't. In order to mitigate this difference, defenders of SV may appeal to a difference between *implicit* and *explicit* representations and hold that the assumptions used in visual computations are only implicitly represented by the system. But this distinction doesn't help.

One way of understanding the implicit/explicit distinction is to take it to denote a difference between representations that are available to the whole cognitive system, including the conscious subject, and representations that are not so available (Fodor, 1983). Surely, the conscious subject does not know (and perhaps wouldn't assent to) the assumptions that the visual system uses to perform its operations. Such assumptions are encoded in proprietary and dedicated representations: this means

that the assumptions are only used within the visual system and their only purpose is to enable visual inferences. Nevertheless, such assumptions are encoded in, or represented by, the system. Thus they constitute a kind of knowledge held by the system.

Another way to clarify the implicit/explicit distinction is to say that the assumptions don't need to be stated as logical rules in propositional format. Palmer says that the assumptions could be "embedded in the pattern of interconnections within a complex neural network" (1999, p. 83). But it is not clear what this proposal amounts to: if being "embedded in the pattern of interconnections" of the neural network means that we should take the pattern to represent the assumptions, then Palmer is conceding that the assumptions need to be encoded *within* the system in some format or other. Presumably, they would need to be so encoded because they have to play a causal role in visual processes.

If, instead, Palmer's proposal is meant to de facto deprive the representation of assumptions of any role in the visual processes (the assumptions embedded in a neural network would lack a syntax and so, arguably, they would also lack causal powers) then we seem to lose track of the sense in which visual systems perform inferences and use knowledge of the world. SV collapses into EV. For, if the assumptions do not need to be represented within the system, then it is not clear how the visual system could follow them in inferring representations. But then talk of inferences would be highly metaphorical: the visual system could, at best, only be *described* as following the assumption of spatial coincidence without actually doing so.

Palmer seems to have in mind the latter worry when, in the passage following the one just quoted, he says:

> Using the term 'inference' to describe such a process may seem to be somewhat metaphorical and thus to undercut the force of the claim that perception works by unconscious inference. But, as we said at the outset, unconscious inference must be at least somewhat metaphorical, since normal inference is quite clearly slow, laborious, and conscious, whereas perception is fast, easy, and unconscious. The important point for present purposes is that perception relies on processes that can be usefully viewed as inferences that require heuristic assumptions. (1999, p. 83)

Although Palmer moves on to talk of visual processing as inferential as if he had fully addressed the problem, the passage highlights an important worry for his view, and for SV more generally. In order to be a substantive claim and not a *mere* metaphor, the inferential account has to hold that mental processes are not just *describable as* inferences (lots of things are), but that they *are* inferences.[5] If the claim is simply that visual systems act *as if* they are performing inferences, or that they can be "usefully viewed" as performing inferences, then the standard account of vision loses it substance. For, the claim that something can be described as inferring is hardly interesting and difficult to disagree with. Lots of things that are not intelligent in the way vision is supposed to be can be described as performing inferences. Take fire alarms: we can describe fire alarms as *wanting* to signal the presence of fire, and doing so whenever they *perceive* smoke because they *know* that smoke is typically

caused by fire. So they *infer* the presence of fire when smoke is present. Would this show that fire alarms perform inferences too and are, after all, intelligent systems? Hardly. Given that we can explain all that the fire alarm does without thinking of it as intelligent, we can continue to suppose that the alarm is just a device designed to signal a certain environmental quantity. Defenders of SV cannot hold a similar position with respect to visual systems. SV is supposed to be a substantive claim about perceptual processes. According to SV, one fundamental difference between cognitive systems, like visual systems, and fire alarms is that cognitive systems literally operate on mental representations in virtue of rules (Palmer, 1999, p. 5); fire alarms, by contrast, do not. Palmer says: "visual perception concerns the acquisition of knowledge. This means that vision is fundamentally a cognitive activity (from the Latin *cognoscere*, meaning to know or to learn) distinct from purely optical processes such as photographic ones" (1999, p. 5).

Cameras, like fire alarms, can be described as performing inferences, but they do not actually do so. As a result, cameras have, according to Palmer, no perceptual capabilities at all, because they "do not know anything about the scenes [they] record" (1999, p. 5). This shows that if, in line with the inferential view, we want to hold that visual systems are intelligent systems, we need to hold that they, in some literal sense, perform inferences (although very fast ones), and that they know about the environment in which they are situated. This means accepting the idea that both features of the environment and assumptions concerning its make-up are represented within the system. EV denies this idea. The system can rely on regularities without representing them. Thus, it does not need to go through a number of representational stages, and can produce representations of objects directly from retinal stimuli. EV is thus a real competitor to SV.

The contrast between SV and EV can be further illustrated if we consider the relation that the visual system holds to the environment. EV requires that we understand the visual system as located in an environment in which it evolved: the system performs its function by being thus located, because it relies on environmental regularities to work as it does. No such location is required in SV, because the resources that the visual system needs in order to work are thought to be *internal* to the system. Assumptions and rules are represented *within* the system. So, although the system will typically be located in an environment, it does not need the environment to function as it does.

Accordingly, while SV presupposes a kind of internalism, EV is a form of externalism, only different from the traditional externalism about content advocated, for example, by Burge (1979): it is rather a kind of process externalism, similar to the one defended by Wilson (2004, chapter 7) and Rowlands (1999, chapter 5), where mental processes are carried out in an environment and can only be understood by making reference to it.[6] This kind of externalism has consequences for how to understand the task of vision scientists. In SV, the assumptions that allow visual computations are thought to be part of the program of the visual system. In EV they describe, instead, environmental constancies. Accordingly, the vision scientist is not out to discover the encoded program of our visual system, but the environmental

conditions on which it relies. In this sense, the embedded approach I favor, is inspired both by Gibson's "ecological optics" (Gibson, 1979) and, more recently, by Rowlands' "environmentalism" (Rowlands, 1999, chapter 5). According to both, part of the task of vision scientists is to discover the complex relation between properties of the environment and properties of the light that hits the retina. Far from discrediting the work done by researchers in the computational or Bayesian paradigms, this externalist approach only reframes their findings as the discovery of the environmental regularities on which visual systems rely.

Having clarified the difference between SV and EV, it is time to turn to the issue of explanatory power. It needs to be shown that EV is able to explain the range of visual phenomena that are accounted for by SV, in particular the phenomenon of multi-stable perception.

3. Multi-Stable Perception without Interpretation

In sections 1 and 2, I presented and contrasted two plausible ways of understanding visual processing. The two ways differ substantially in the amount of encoded resources that they ascribe to the visual system. The Embedded View is radically more parsimonious in ascribing knowledge of the world to the system. This may be taken to be an advantage of the view; but EV can be preferred to its more traditional alternative only if we show that it is as explanatory powerful as the Standard View. This section is dedicated to this task.

To introduce how the EV explains multi-stable perception, notice, first, that EV is well poised to explain the array of visual phenomena that proponents of SV mention in support of their position. Visual stability, misperceptions and illusions are all explained, by SV, by making reference to information about physical facts possessed by the system. EV differs only in noticing that the visual system can rely on the facts without possessing information about them. EV does not do without talk of assumptions, constraints and regularities: it simply conceives of the visual system as relying on them, rather than representing them. In so doing, the account avails itself to many of the resources of the inferential view, without taking on some of its dubious commitments, in particular the idea that visual systems possess knowledge of a striking number of physical facts.

Take the problem of visual stability first: the problem consists in understanding how visual systems derive a single percept from a retinal stimulation that is compatible with a number of distal causes. According to SV, perceptual systems use some stored knowledge of the world to reduce the possible interpretations of the retinal data. By contrast, EV presumes that the visual system, having evolved in a specific environment, relies on its regularities to produce unique representations. The environment, rather than the visual system, constraints the visual representations produced given the stimulus.

To clarify what this means, consider the following, somewhat imaginary, example: suppose that a given retinal pattern is, in our environment, most often caused by

edges and, on sporadic occasions, by cracks, perhaps because cracks are less common than edges. We can think of the visual system as somehow knowing this fact, thus producing representations of edges given the input, through an inferential process. This allows the system to get it right on most occasions because edges are more commonly associated with the given retinal pattern. But it is just as reasonable to suppose that the visual system is built, given the retinal stimulus, to represent edges rather than cracks *directly*. Given that edges are the typical causes of the stimulus, it would be surprising if it were otherwise. The environment in which the system evolved is one where, given the retinal pattern, the probability of getting it right by representing edges is higher than the probability of getting it right by representing cracks. If we suppose, like SV does, that getting it right has some evolutionary value, we may also suppose that the visual system is wired to produce representations of edges rather than representations of cracks. The system *relies* on the fact that edges are the typical causes of the given stimulus, and it represents edges without using any stored knowledge of that fact. In the event in which cracks are present instead, the system misrepresents the environment.

Notice that the tendency to suppose that the visual system needs to possess knowledge of the world in order to solve the underdetermination problem is made plausible only if we think of vision as an isolated system. We can resist this tendency if we think of it as embedded in an environment in which it evolved. The complexity that SV ascribes to the visual system is offloaded to the environment.

Now consider the inverse problem to the one of stable perception: we need to understand how vision can sometimes be unstable. Here, SV appeals again to the notion of interpretation, but it is hardly ever spelled out how this notion helps. The recurring idea is that visual stimuli can be interpreted in multiple ways. This supposedly means that different assumptions can be used in computing the stimulus, and so different representations can be inferred from it. The visual system might, for example, sometimes assume that a discontinuity in light intensity is caused by an edge, and sometimes assume that it is caused by a crack, thus producing different perceptions of the same environmental object.

It is a bit harder to see what the assumptions would be in cases like the duck-rabbit, or the vase-face figures. So far, we have mentioned assumptions about edges and rigid objects, but these can hardly be the whole story of how we can see, say, a figure first as duck-shaped and then as rabbit-shaped. Perhaps we need to ascribe to visual systems not only knowledge of edges and rigid objects, but also knowledge of ducks and rabbits. This would be somewhat implausible, but, luckily, we don't need to do it. This is because we can think of whatever assumptions the visual system needs, not as encoded within it, but as relied upon. Just like the visual system can make use of different encoded assumptions, it can also rely on different environmental facts. Thus, when presented with a sharp discontinuity in light intensity it can sometimes rely on the fact that edges typically cause it, and sometimes rely on the fact that cracks sporadically cause it. Similarly, it can represent the

presence of something duck-shaped or of something rabbit-shaped when doing so accords with the retinal stimulus. The system does not interpret the stimulus as produced by something rabbit-shaped: it rather relies on the fact that the stimulus is typically (or often) caused by something that looks like a rabbit.

For an analogy, consider, again, a fire alarm. Suppose that we built a fire-alarm to signal not only the presence of fire but also the presence of running cars, perhaps by emitting different sounds. There are at least two ways in which we could build such an alarm. We start by noticing that smoke can be caused both by fire and by running cars. In this sense, smoke is an underdetermined or ambiguous input because it can be caused by different environmental causes. One way to build our alarm is to build a system that represents the presence of smoke and then uses one of two assumptions: that smoke is caused by fire, and that smoke is caused by running cars. The system typically uses only one of these assumptions, inferring, say, the presence of fire. But sometimes it uses the other, inferring the presence of a car, and on some occasions it can switch back and forth from using one to using the other. When this switch happens, the fire alarm produces two different signals and alternates from one to the other. This would be an inferential fire alarm that is capable of ''multi-stable signaling.'' Alternatively, we could build an embedded fire alarm, one that relied on a couple of environmental facts: that smoke is caused by fire, and that smoke is caused by running cars. The alarm typically relies on just one of these facts, signaling the presence of fire. But sometimes it relies on the other, signaling the presence of a car, and on some occasions the alarm can switch back and forth from relying on one fact to relying on the other. This would be an embedded fire alarm that is capable of multi-stable signaling. Notice that this type of fire alarm requires much less internal complexity, in particular less programming.

Admittedly, this preliminary explanation of embedded multi-stable perception/signaling leaves much to be desired. One thing that is left unexplained is why the visual system sometimes relies on different facts given the same stimulus, to produce different percepts. In other words, we still need to explain what conditions prompt the shift given that perception is typically stable. A detailed explanation of this fact, compatible with an embedded framework is outside the scope of this paper. I'll briefly outline such explanation below, but let me point out, first, that SV has the same problem: SV, just like EV, has to explain why the visual system sometimes uses different assumptions to interpret the same visual stimulus. That is, it has to explain what prompts the shift given that perception is typically stable. But then EV and SV are equivalent both in explanatory power and in what they leave unexplained: appealing to an interpretation, by itself, is not sufficient to account for why we sometimes get two percepts rather than the usual one by looking at a single object or figure. Accordingly, if what I have argued so far is right, an inference or an interpretation is neither necessary nor sufficient to explain aspect shifts. An embedded and more parsimonious account will do. In the next section, I offer a very preliminary picture of the kind of theory of multi-stability we get by accepting the embedded framework.

4. Multi-Stable Perception as Visual Search

Proponents of SV tend to appeal to the creative and interpretative nature of our visual system in order to explain our ability to see objects and figures in multiple ways (Palmer, 1999, p. 10). In the last section, I argued that such appeal is neither necessary nor sufficient. It is not necessary because we can think of the visual system as a simple device that has been built to produce different signals, and in particular different representations, by relying on different environmental facts. It is not sufficient because it leaves unexplained what prompts the shift given that perception is typically stable.

Here, I propose to replace the idea that multi-stability involves creativity with the idea that it involves visual curiosity. This idea rids us of the need to appeal to an interpretation, and it gestures towards an explanation of the occasional instability of visual perception. The occurrence of shifts is guided by visual search and, in particular, by two factors: the ambiguity of the stimulus *relative* to our specific environment, and the role of attention. Appeal to these two features highlights again the importance of collocating the visual system in the world, and understanding it by making reference to such world.

The first idea is, roughly, that certain light patterns, although compatible with many distal causes, have, in our environment, one typical cause; that is, they are more likely to be caused by one type of entity. When this is the case, perception is usually stable, because the stimulus, although ambiguous in principle, is not ambiguous in our specific environment. Other light patterns, by contrast, have more than one typical cause: they are as likely to be caused by something, say, duck-shaped, as they are to be caused by something rabbit-shaped. When this is the case perception is predictably unstable. In this framework, stability and instability are explained, at least in part, by making reference to features of the visual stimulus. That is, they are explained by making reference to what is *external* to the visual system (properties of light) rather than internal to it. Subtle properties of the pattern of light make it more likely that one environmental entity rather than another caused it (or that two or more entities are as likely to cause it). Similarly, we may suppose that subtle properties of smoke make it more likely that fire rather than a car engine caused it (or that fire and a car engine are as likely to cause it).

So, the first step in explaining the instability of perception consists in appealing to the contingent instability of the stimulus. This accounts for why certain objects or figures are more likely than others to give rise to multi-stable perception: they project light on the retina that is ambiguous relative to our environment. In simpler words, they are ambiguous. Other objects and figures project a pattern of light that, although ambiguous or underdetermined in principle, is not underdetermined in practice, that is, relative to our environment. In this sense, they are not ambiguous.

The other step consists in recognizing the role of attention in aspect-shifts. Multi-stable perception is a phenomenon that has a temporal dimension: one *first* sees a figure as duck-shaped and *then* as rabbit-shaped. Recognizing this temporal dimension, allows us to see aspect-shifts as plausibly guided by visual exploration,

in particular by paying attention to different parts of a figure or object. Paying attention to some features may reveal properties of the light that are more likely associated with one environmental object, while paying attention to others may uncover properties of the light that are more likely associated with a different object. Disambiguation is then achieved by performing visual search, and this is fully compatible with an embedded account of visual processing: by paying attention to certain parts of a figure the visual system is exposed to features of light that are likely caused by something, say, duck-shaped. The system relies on this fact to produce a representation of something duck-shaped. Alternatively, by paying attention to other parts of the figure, the system is exposed to features of light that are likely caused by something rabbit-shaped. And, again, by relying on this fact the system produces a representation of something rabbit-shaped.

What emerges is a picture of multi-stability in which the visual system is out to find what is already present to the senses, rather than putting it there through an inference or a construction. Contrary to SV, nothing particularly creative is involved in seeing the world in different ways: shifts are prompted by curiosity, rather than by inventiveness. In the externalist spirit: we see the world in different ways because of how the world is, not because of how we construct it to be.

A body of empirical evidence confirms the importance of visual search in multi-stability. Attentional mechanisms play a significant role in shifts. In adults, the part of a reversible figure that the observer focuses on has been shown to determine which percept the viewer experiences (Chastain & Burnham, 1975). Neurological studies confirm the stimulation of pre-frontal areas associated with attention during reversals (Britz, Landis, & Michel, 2009; Kleinschmidt, Buchel, Zeki, & Frackowiak, 1998; Leopold & Logothetis, 1999; Nakatani & van Leeuwen, 2006; Sterzer & Kleinschmidt, 2007; Tong, Meng, & Blake, 2006) and patients with unilateral frontal brain damage have greater difficulty in shifting than normal control subjects (Ricci & Blundo, 1990).

Studies on ocular activity during reversals further support this picture. Shifting is a phenomenon that extends in time and that involves a number of ocular events. In particular, saccades prior to, and during, reversals are positively associated with the process. They increase in the period leading to a shift (Ito et al., 2003), and in conditions where fixation instructions restrict them perceptual switching rate is considerably reduced (Glen, 1940; Toppino, 2003). Since saccades are closely associated with shifts in spatial attention (Slotnick & Yantis, 2005) this again implicates attentional mechanisms in the switching process.

Research on bilingual children further supports this view: bilingual children at 6 years of age are more likely to experience reversals than their monolingual peers (Bialystok & Shapero, 2005). This result can be attributed to the fact that bilingual children develop control over selective attention earlier than monolingual children because they have to control two active language systems (Bialystok, 2001; Bialystok & Martin, 2004). Further, children younger than 5 are usually unable to reverse ambiguous figures spontaneously because pre-frontal cortex develops relatively late in humans reaching maturation only during adolescence (Diamond, 2002).[7]

More needs to be said to spell out and assess the merits of this positive account of multi-stable perception, and this is not the place to do so. One thing, however, should be clear from the previous discussion: SV and the EV are explanatorily on a par. EV is just as capable of explaining aspect shifts as SV. By thinking of the visual system as relying on different environmental facts, we can explain how we get multiple percepts from looking at the same figure or object. But then the supposed explanatory superiority of SV fades.

5. Conclusion

I have argued that a non-inferential and direct account of vision is capable of explaining multi-stable perception. The embedded visual system can exploit different environmental regularities without representing them, thus producing different representations of objects and scenes without performing inferences. The primary difference between the embedded account I favor and standard views of visual processing is the amount of representational resources ascribed to the visual system. That EV is more parsimonious can be counted in its favor. If, as I have argued, the proclaimed explanatory superiority of SV is only apparent, then we have the beginning of an argument for an embedded understanding of visual processing.

Acknowledgments

Thanks to Dorit Bar-On, Casey O Callaghan, Bill Lycan, Ram Neta, Jesse Prinz, Dylan Sabo and an anonymous referee for inspiration and advice.

Notes

[1] I take this characterization of representation to be in line with that of proponents of the Standard View (Marr, 1982, p. 80; Segal, 1989, p. 194).

[2] Sabo (unpublished manuscript) develops a similar and illuminating account of our mechanism of concept acquisition. The idea of relying on environmental regularities can also be found in the work of Zenon Pylyshyn (1999) and Robert Wilson (2004, p. 163). Pylyshyn talks about the visual system ''embodying'' the assumptions rather than representing them; Wilson talks about ''exploiting'' the assumptions or using exploitative representations of them. I prefer the jargon of ''relying on'' regularities in order to make clear that the assumptions are not encoded anywhere within the visual system.

[3] This result reflects a distinction between what are sometimes called ''explanatory reasons'' and ''justificatory reasons'' (Dretske, 2006, pp. 28–29). Explanatory reasons are facts that explain, or help to explain, why something happens: they are the reasons *why* something happens. Justificatory reasons, on the other hand, are given by the way in which facts are represented to be: they are reasons *for* something to happen. Justificatory reasons are also explanatory, but they explain by making reference to the way things are represented to be. In the fire alarm case, the fact that smoke particles deflect the direction of light is one of the explanatory, not justificatory, reasons for what the alarm does.

[4] For similar points concerning the difference between representations and states that are merely sensitive to environmental quantities, see Sabo (unpublished manuscript) and Burge (2010, chapters 2 and 8).

[5] Similarly, proponents of the view that visual processes are computational hold that visual processes *are* computations, not just that they are *describable as* computations. The latter approach would favor a kind of pancomputationalism that is actually in tension with the Computational Theory of Mind (Piccinini, 2007).

[6] I leave it open here whether the visual system is not only embedded in the environment, but it also forms a coupled system with the environment and can be said to extend beyond the skin's boundaries (Clark, 1997; Wilson, 2004). There is considerable debate concerning the plausibility of the extended view (Adams & Aizawa, 2009; Rupert, 2004) and I do not have space to approach the debate here.

[7] This body of evidence indicates that the capacity to direct and hold attention to certain features of a scene is positively correlated with reversals. More work needs to be done to establish the exact role of attention in shifting, in particular in order to see if attentional shifts are a necessary condition for reversing. Some studies suggest that attention plays a more modest role in binocular rivalry than in multi-stability in the presence of ambiguous images (Meng & Tong, 2004; van Ee, Noest, Brascamp, & van den Berg, 2006). This, however, is relatively unsurprising. In rivalry two different stimuli are presented to each eye, and rather than fusing them into a single image, we experience a shift between perceiving one and perceiving the other. Although this phenomenon is similar to perceptual shifts in the presence of ambiguous figures, it is importantly "stimulus-driven." The *two* images plausibly drive the shift without needing to perform visual search. By contrast, we should expect attention to play a bigger role in perceptual multi-stability where the *same* stimulus is perceived in two or more ways. This may still not imply that attention is a necessary condition for this kind of phenomenon, but the correlational evidence is certainly significant.

References

Adams, F., & Aizawa, K. (2009). Why the mind is still in the head. In P. Robbins & M. Aydede (Eds.), *The cambridge handbook of situated cognition* (pp. 78–95). Cambridge: Cambridge University Press.

Bialystok, E., & Martin, M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*, 7, 325–339.

Bialystok, E., & Shapero, D. (2005). Ambiguous benefits: The effect of bilingualism on reversing ambiguous figures. *Developmental Science*, *8*, 595–604.

Britz, J., Landis, T., & Michel, C. M. (2009). Right parietal brain activity precedes perceptual alternation of bistable stimuli. *Cerebral Cortex*, *19*, 55–65.

Burge, T. (1979). Individualism and the mental. In P. A. French, T. E. Vehling, & H. K. Wettstein (Eds.), *Midwest studies in philosophy* (Vol. 4, pp. 73–121). Minneapolis, MN: University of Minnesota Press.

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

Chastain, G., & Burnham, C. A. (1975). The first glimpse determines the perception of an ambiguous figure. *Perception and Psychophysics*, *17*, 221–224.

Churchland, P. (1989). Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor. In his *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 255–279). Cambridge, MA: MIT Press.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

Diamond, A. (2002). Normal development of pre-frontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. Stuss & R. Knight (Eds.), *Principles of frontal lobe functioning* (pp. 466–503). New York: Oxford University Press.

Dretske, F. (2006). Perception without awareness. In T. Gendler & J. Hawthorne (Eds.), *Perceptual experience* (pp. 147–180). Oxford: Claredon Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1988). A reply to Churchland's ''Perceptual plasticity and theoretical neutrality''. *Philosophy of Science*, *55*, 188–198.

Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's ''ecological approach''. *Cognition*, *9*, 139–196.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Glen, J. S. (1940). Ocular movements in reversibility of perspective. *The Journal of General Psychology*, *23*, 243–281.

Gregory, R. L. (1970). *The intelligent eye*. New York: McGraw-Hill.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*, 687–701.

Ito, J., Nikolaev, A. R., Luman, M., Aukes, M. F., Nakatani, C., & van Leeuwen, C. (2003). Perceptual switching, eye movements, and the bus paradox. *Perception*, *32*, 681–698.

Kleinschmidt, A., Buchel, C., Zeki, S., & Frackowiak, R. S. (1998). Human brain activity during spontaneously reversing perception of ambiguous figures. *Proceedings of the Royal Society B: Biological Sciences*, *265*, 2427–2433.

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Science*, *3*, 254–264.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman and Company.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society B: Biological Sciences*, *207*, 187–217.

Meng, M., & Tong, F. (2004). Can kattentong selectielid bias postale perceptiën? Differences between binocular rivalry and ambiguous figures. *Journal of Vision*, *4*, 539–551.

Nakatani, H., & van Leeuwen, C. (2006). Transient synchrony of distant brain areas and perceptual switching in ambiguous figures. *Biological Cybernetics*, *94*, 445–457.

Orlandi, N. (unpublished manuscript). Embedded seeing: Vision in the natural world.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Piccinini, G. (2007). Computational modeling vs. computational explanation: Is everything a turing machine, and does it matter to the philosophy of mind? *Australasian Journal of Philosophy*, *85*, 93–115.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341–423.

Ricci, C., & Blundo, C. (1990). Perception of ambiguous figures after focal brain lesion. *Neuropsychologia*, *28*, 1163–1173.

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

Rowlands, M. (1999). *The body in mind*. Cambridge: Cambridge University Press.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, *101*, 389–428.

Sabo, W. (unpublished manuscript). Concept acquisition without representational mediation.

Segal, G. (1989). Seeing what is not there. *The Philosophical Review*, *98*, 189–214.

Slotnick, S. D., & Yantis, S. (2005). Common neural substrates for the control and effects of visual attention and perceptual bistability. *Cognitive Brain Research*, *24*, 97–108.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*, 29–56.

Sterzer, P., & Kleinschmidt, A. (2007). A neural basis for inference in perceptual ambiguity. *Proceedings of the National Academy of Science, USA*, *104*, 323–328.

Tong, F., Meng, M., & Blake, R. (2006). Neural bases of binocular rivalry. *Trends in Cognitive Sciences*, *10*, 502–511.

Toppino, T. C. (2003). Reversible-figure perception: Mechanisms of intentional control. *Perception & Psychophysics*, *65*, 1285–1295.

Ulman, S. (1979). *The interpretation of visual motion*. Cambridge, MA: MIT Press.

van Ee, R. (2003). Bayesian modeling of cue interaction: Bistability in stereoscopic slant perception. *Journal of the Optical Society of America B*, *20*, 1398–1406.

van Ee, R., Noest, A. J., Brascamp, J. W., & van den Berg, A.V. (2006). Attentional control over either of the two competing percepts of ambiguous stimuli revealed by a two-parameter analysis: Means do not make the difference. *Vision Research*, *46*, 3129–3141.

Wilson, R. (2004). *Boundaries of the mind*. Cambridge: Cambridge University Press.