Indexing the World? Visual Tracking, Modularity,

and the Perception–Cognition Interface

SANTIAGO ECHEVERRI

ABSTRACT

Research in vision science, developmental psychology, and the foundations of cognitive science has led some theorists to posit referential mechanisms similar to indices. This hypothesis has been framed within a Fodorian conception of the early vision module. The paper shows that this conception is mistaken, for it cannot handle the 'interface problem,' roughly, how indexing mechanisms relate to higher cognition and conceptual thought. As a result, I reject the inaccessibility of early vision to higher cognition and make some constructive remarks on the perception–cognition interface.

Research in vision science, developmental psychology, and the foundations of cognitive science has led a number of theorists to posit primitive perceptual relations to objects similar to indices (Campbell [2002]; Pylyshyn [2003], [2007]; Raftopoulos [2009]; Dickie [2010], [2011]). In his influential work on this topic, Zenon Pylyshyn framed this hypothesis within a model of mental architecture that combines a Fodorian view of the early vision module with a symbolic model of cognition. This paper shows that, although empirical findings in favor of tracking mechanisms are robust, they are not modular in Fodor's sense. In defense of this claim, it is shown that this picture of modularity cannot handle the 'interface problem,' roughly, how indices relate to higher cognition and conceptual thought.

The paper falls into four sections. Section 1 summarizes some arguments for the existence of indices. Section 2 explains how one could integrate those indices within a model of mental architecture that includes perceptual modules of a Fodorian sort. Section 3 offers a characterization of the interface problem, and argues that the Fodorian conception of modules cannot handle two aspects thereof: top-down forms of attention and the gap between visual and ordinary objects. In section 4, I make some suggestions on how the Fodorian conception of modularity could be revised to provide an adequate account of the perception–cognition interface.

1. The Case for Visual Indices

This section summarizes some influential arguments for the claim that visual systems have mechanisms analogous to indices. To this end, I first introduce some general assumptions, and then present the theoretical and empirical arguments that motivate this view.

1.1. Preliminary Assumptions

When one argues that perception involves relations to objects, one has to be clear on what is meant by the notion of an object. For the purposes of this paper, I will define objects as entities *governed* by a number of principles (Bermúdez [2003]). I call these principles 'object constraints.' Object constraints are minimal conditions that portions of reality must satisfy in order to be parsed as objects. Although the notion of an object constraint deserves a detailed analysis, we can get an intuitive grasp thereof by contrasting it with the notion of a property. Properties like BLUENESS or SQUARENESS are usually thought to be had by objects. This suggests that they might be detached from the object, and the object would remain numerically the same. By contrast, object constraints seem to be constitutive of objects. If a constraint is violated, one may move from counting a scene as having one object to having two objects or no object at all. Here are two influential constraints:

> <u>Continuity</u>: Humans attend to and keep track of entities that seem to follow a continuous path through space and time.

Persistence: Humans understand objects as the same when changes in properties (such as color or shape) occur, and after brief periods of occlusion. (Spelke [1990]; Carey [2009])

Although objects such as chairs, trees or rocks respect these constraints, satisfying them may be insufficient for an entity to count as an object in the *ordinary* sense of the term. After all, entities we do not usually count as ordinary objects also satisfy these conditions. Examples include dots on screens, vertices of geometrical shapes, ripples on water, patches of reflected light, and so on. (Dickie [2010], p. 221) That is why they have been called 'visual objects.' (Carey [2009]; Pylyshyn [2003], [2007]; Scholl [2007]) Arguably, the ability to single out visual objects is a precondition of the ability to refer to ordinary objects. Hereafter, when I speak of objects, I mean entities that satisfy these (and probably other) constraints.

Since object constraints are respected by human visual systems, there is a sense in which they demonstrate the existence of *distinctive* tracking mechanisms. But some theorists have gone farther, and held that human visual systems exploit indices to keep track of objects. In what follows, I summarize some considerations for this view: (1.2) a series of transcendental arguments and (1.3) some findings on visual attention and the tracking of moving targets.

1.2. Transcendental Arguments

A host of considerations for the existence of indices originates from Pylyshyn's ([2003],

[2007]) influential work on the foundations of cognitive science. One of his most interesting arguments derives from his commitment to a symbolic picture of cognition. On this view, cognitive processes are transformations of symbols in a language of thought (LOT). This raises a question: How are those symbols connected with the external world? Clearly, some of the symbols are predicates of the form *F(x)*, *G(x)*, *H(x)*, and so on. Suppose now one introduces these predicates to explain the detection of visual properties. Then, there must be a mechanism that binds these predicates to objects.

Such remarks provide the materials for a series of transcendental arguments. To a first approximation, a transcendental argument takes some initial conditions as given, and proceeds to establish the existence of a fact that must obtain if those conditions are to be intelligible (Stern [2011]). In the present case, if some perceptual and cognitive processes are defined over descriptive representations, there must be a mechanism that binds them to objects. If successful, the arguments establish a conditional conclusion: *If* some mental processes operate on descriptions, there *must* be direct relations to objects.

Here is the most general of those arguments. Suppose one tries to explain how the predicate *F(x)* is anchored to an individual *a*. In order to yield a representation of the form *Fa*, the process could not rely on a further predicate *I(x)*. After all, if the selection of an object required a prior application of the predicate *I(x)*, we would get *F(I(x))*. If the only way of referring to a particular consisted in encoding a property, we would be launched in a regress. Since there is no such regress—because we do refer to individuals—, there must be non-predicative ways of referring to objects.

Pylyshyn’s argument is structurally similar to those a number of philosophers made a few years ago, when they invoked colorful metaphors to convey the basic idea of *de re* thought. Here is Kent Bach:

> If *all* thoughts about things could only be descriptive, your total conception of the world would be merely qualitative. You would never be related to anything in particular. Thinking of something would never be the case of having it 'in mind,' as we say colloquially, or as some philosophers have said, of being '*en rapport,'* in 'cognitive contact,' or 'epistemically intimate' with it. But picturesque phrases aside, just what is this special relation? Whatever it is, it is different from that involved in thinking of something under a description. […] Since the object of a descriptive thought is determined SATISFACTIONALLY, the fact that the thought is of that object does not require any connection between thought and object. However, the object of a *de re* thought is determined RELATIONALLY. For something to be the object of a *de re* thought, it must stand in a certain kind of relation to that very thought. (Bach [1987], p. 12; see also Campbell [2002]; Pylyshyn [2007], p. 8, p. 12; and Recanati [2012], p. 19)

If cognitive processes are symbolic, and some symbols are predicates, there must be non-descriptive modes of reference that bind those predicates to objects in the world. In other words, not all reference can be determined by the satisfaction of properties encoded in LOT.

More specific arguments for the existence of non-satisfactional mechanisms of reference arise from studies on visual attention and visual reasoning. An example of these arguments comes from Ullman's ([1984]) research on visual routines (Figure 1).

According to Ullman, the detection of some visual properties—like BEING INSIDE A CLOSED CONTOUR or BEING COLLINEAR—requires the execution of serially applied operations by which the visual system scans the scene. In order to implement these serial processes in a symbolic model, there must be a mechanism that explains how the system keeps track of the elements it has already 'visited,' and those it 'visits' afterwards. According to Pylyshyn, this requires a mechanism to refer to the visual elements *as* individuals without encoding their visual properties. This mechanism should operate in parallel, allowing the mind to maintain 'referential contact' with a number of different items at the same time (Scholl [2009], p. 53). After all, the purpose of visual routines is to evaluate whether two items, say *z* and *x* in Figure 1a, satisfy a property (Pylyshyn [2003], p. 207). Other researchers inferred from similar studies that visual routines require mechanisms of mental tagging (Ballard *et al.* [1997]; Olivers, *et al.* [1999]); here 'tagging' should be understood as a non-descriptive mode of reference.

[Insert Figure 1]

Visual routines provide an example of a fundamental issue of vision science: the correspondence problem. If visual representations are constructed over a series of glances, how does the system establish correspondences between objects at different times, if those objects present different properties at different stages of the construction process? According to Pylyshyn, this problem could not be solved if the system *merely* had descriptive representations (Pylyshyn [2003], pp. 146–50, p. 205, pp. 245–6). This leads him to introduce visual indices:

> Unless at some point one can think the equivalent of "*This* has property *P*," one cannot refer to a particular object token in a way that will allow it to be bound to the arguments of a visual predicate or to serve as the basis for action (e.g., to point toward or grasp the token object in question). (Pylyshyn [2003], p. 254; see also his [2007], p. 95)

These transcendental arguments yield a feature of binding mechanisms: they are non-descriptive modes of reference. Yet, they leave two crucial questions open: What is the nature of these non-descriptive modes of reference? How do these non-descriptive modes of reference block the regress? Let us examine each question in turn.

Prima facie, non-descriptive modes of reference might be implemented by proper names, demonstratives or even complex descriptions of the form 'the *F(…x…)*,' where *'…x…'* is not purely qualitative but contains at least one singular expression. Nevertheless, Pylyshyn—as many other theorists—proposes that the non-descriptive mechanisms of reference are similar to demonstratives. There are two reasons for this choice: First, whereas qualitative descriptions and proper names can refer to items *in absentia*, the most central uses of demonstratives require that the observer stand in a spatiotemporal relation to the referent (Pylyshyn [2007], pp. 16–7, p. 82, p. 122).[1] Second, there is a fundamental contrast between demonstratives and proper names: whereas the latter keep their referent constant, the former can acquire different referents in different contexts. The latter property is often called 'shiftability.' As we shall see later, this feature gets some support from experimental evidence. Thus, compared with names and qualitative descriptions, demonstratives offer a more plausible model of visual reference.[2]

The second issue is more delicate, though. How do visual indices block the regress? Pylyshyn's reasoning, as that of many people working on *de re* thought, has the form of a 'what-else' argument: If the reference of indices is not satisfactional, what else but a direct relation can determine the referent? And what else but a non-cognitive and purely causal process could explain it? Pylyshyn's proposal is to avoid the regress by *stipulating* that the referent of indices is determined in a *direct*, *non-cognitive*, and *causal* manner. As he points out: 'Sooner or later concepts must be grounded in a primitive causal connection between thoughts and things. The project of grounding concepts […] in perception remains an essential requirement if we are to avoid an infinite regress.' (Pylyshyn [2001], p. 154; see also his [2007], p. 17, p. 33, p. 57)

In section 4, I will take issue with this last step of Pylyshyn's argument. For the time being, it is worth stressing that arguments of this sort play a heuristic role in empirical science. Although they do not provide a neat-and-tidy logical entailment to the desired conclusion, they can be seen as making a plausible case for the existence of some facts, for example indices that refer in a non-satisfactional manner. This is precisely what occurred to Pylyshyn and his co-workers. The transcendental arguments led them to design new experimental setups that generated new evidence and to reinterpret other findings in a way that is consistent with the hypothesis of indices.

1.3. Evidence from Vision Science

There is also empirical evidence for direct links to objects. The first data come from a number of studies that demonstrate that the visual system can determine the cardinality of a few elements in parallel. This robust phenomenon is known as *subitizing* (Intriligator and Cavanagh [2001]). If one is presented with a small number of items (up to four or five), one can provide a fast and accurate response to the question: How many items are there? Speed, accuracy, and confidence decrease conditional on the number of items and their spatial distribution (Figure 1d). On Pylyshyn's view, this lends support to the claim that humans have perceptual means of singling out between four and five items in a direct way (Pylyshyn [2003], pp. 174–6).

[Insert Figure 2]

Pylyshyn and his co-workers also created an experimental paradigm to test the seminal transcendental arguments: multiple-object-tracking (MOT). In a typical MOT experiment, subjects are presented with various qualitatively identical items on a computer screen (Figure 2). Some of them are flashed to indicate their status as targets. After that, the objects start moving randomly for some time (usually ten seconds). At the end, observers are asked to indicate the original targets. These studies show that humans can keep track of four or five moving targets in parallel, confirming the results on subitizing. Some variations of this paradigm lend support to the further idea that the system has a direct mode of reference. Observers can track individuals even though they change some of their visual properties, like shape and color. In some cases, subjects do not seem to detect those changes. If one assumes that verbal reports reliably indicate whether some properties are encoded, these findings would show that some visual properties are not recruited for visual tracking.[3] Since the targets are following spatiotemporal paths on the screen, however, MOT experiments suggest that visual tracking is governed by a principle of continuity (section 1.1).

Further studies have shown that tracking is governed by other constraints, like persistence (Scholl and Pylyshyn [1999]; for a review, see Scholl [2007]). In some experiments, observers are asked to keep track of objects in spaces filled with virtual occluders. These experiments are based on the plausible idea that objects do not pop into or out of existence, though they do frequently pop into and out of sight. Two conditions have been tested. In some cases, the targets disappear by gradually 'imploding' and later 'exploding.' In others, they disappear in a way that indicates the presence of occluding surfaces, by accreting and deleting along a fixed contour (Figure 3). It has been found that subjects' tracking abilities are not impaired in the latter case. Although subjects know that the occluders are not real (but virtual), this background knowledge plays no role in their ability to track objects, suggesting that the mechanism is cognitively impenetrable (more on this notion in section 2).

[Insert Figure 3]

These findings refine the conclusions of the transcendental arguments. Recall that, if cognitive processes are symbolic, there must be non-descriptive mechanisms of direct reference. Pylyshyn hypothesizes that those mechanisms are purely causal, which means that they are taken as primitive. Empirical studies on subitizing and MOT introduce further tweaks on the initial hypothesis. Humans have mechanisms that enable them to keep track of up to four or five targets, the mechanisms operate in parallel, and are governed by object constraints, like continuity and persistence. Recall also that the transcendental arguments left open the question whether the mechanisms were realized by analogues of proper names, descriptions or demonstratives. The experimental evidence suggests that humans have between four and five tracking mechanisms that can be assigned to different items, depending on the causal relation that obtains between the system and those items. Thus, it seems appropriate to describe them as *shiftable* indices that exploit some spatiotemporal relations to acquire a referent.

With these results in place, it seems natural to try to ground *conceptual* thought about ordinary objects in these primitive tracking mechanisms.[4] Despite the plausibility of this project, however, there is no direct path from the experimental findings on visual tracking to substantial conclusions on the underlying cognitive structures. Although empirical evidence enables us to state some *predictive* principles that govern visual tracking, it remains silent on how these principles are realized. Given that a model of mental architecture is not the sort of thing one can establish through demonstrative argument, it is reasonable to examine the available proposals for their coherence and capacity to encompass the relevant findings. I propose to focus on the most detailed proposal: Pylyshyn's contention that indices are implemented in the early vision system, which he conceives as encapsulated and inaccessible to higher cognition. Yet, the present discussion will be of interest to any theorist who is attracted either by the existence of perceptual demonstratives or by the modularity of some perceptual systems.[5]

2. Visual Indices, Object Files, and Fodorian Modularity

Most available views of perceptual reference exploit the notion of an *object file*. The phrase 'object file' was coined to interpret a series of studies on the binding of properties (Treisman & Gelade [1980]; Kahneman *et al.* [1992]; Treisman [1998]). The goal of this research is to explain how the visual system binds perceptual features together. This is a version of what Jackson ([1977]) terms the 'multiple property problem': when presented with a number of features at the same time, the system must determine which must be bound together. If one is presented with two letters, e.g. a blue 'F' and a red 'G,' it must bind the property BLUENESS with the property of BEING AN 'F' and the property REDNESS with that of BEING A 'G.' Treisman and her co-workers found that, under some experimental conditions, subjects can make mistakes in the binding task. This led her to hypothesize that spatial location might play a crucial role in binding, and to introduce the notion of an object file to explain how spatial location contributes to solving the binding problem.

For reasons of space, I shall not discuss the details of Treisman's account (but see Clark [2000] and Campbell [2002] for philosophically oriented discussions thereof). The important point is that object files are depositories of information in working memory. When one perceives an object, one forms a temporary mental structure dedicated to retrieving information about it. This idea has been used to interpret the results of MOT experiments. These findings lend support to the claim that human adults can form up to four or five object files. A feature is bound to an object if it is fed into the object file that was originally opened for that object. The differences between various theories lie in the way they explain the relation of a file to the object. Given the transcendental arguments presented above (section 1.2), the relation is explained by the activation of indices.

How do indices fit in a broader model of mental architecture? Pylyshyn's proposal is based on two hypotheses:

(H1) The function of indexing mechanisms is to individuate objects before they are conceptually recognized.

(H2) Visual indices are realized by modular systems in Fodor's sense, i.e. they are encapsulated and inaccessible to higher cognition.[6]

In what follows, I spell out these hypotheses. In the next section, I explain why they face what I call the 'interface problem.'

Theorists like Treisman and Pylyshyn use the notion of an object file to convey the idea that any recognition of an object must be preceded by the operation of a mechanism responsible for *individuating* it. Individuation consists in segregating a portion of the world in conformity with the object constraints (Pylyshyn [2003], pp. 134–5, [2007], p. 206). This process gives rise to object files, i.e. buffers that are connected with concepts stored in long-term memory. Recognition is a process in which the system matches object files with stored models or concepts. (Treisman [1998], p. 1297) The latter not only represent visual properties but also high-level properties like BEING A PREDATOR, BEING A GLASS or BEING A COMPUTER (Dickie [2010], [2011]; Fodor [2008], pp. 92–ff.; Recanati [2012]).[7]

The second hypothesis requires more stage setting. According to Pylyshyn's influential characterization, a principle or mechanism is *cognitively impenetrable* if it cannot be *rationally* modified by the cognitive states of the system (Pylyshyn [2007], p. 164). Object constraints are cognitively impenetrable in this sense. Subjects know that there is no room for occlusion in a computer screen. Nevertheless, they cannot avoid perceiving some two-dimensional targets as being momentarily occluded by others.

The modularist explanation of this phenomenon says that, if a mechanism is *cognitively impenetrable,* it is implemented in an *informationally encapsulated* module (Fodor [1983], [1984], [2000]; Pylyshyn [1984], [1999], [2003]). A module can be defined as a dissociable functional component (Carruthers [2006], p. 2). To a first approximation, the notion of encapsulation explains impenetrability by imposing some restrictions on the information flow between the impenetrable module and cognitive systems. I cannot evaluate here the general idea of a module (for criticism, see Prinz [2006] and Samuels [2006]). I will simply assume that encapsulation provides a plausible (and widespread) explanation of impenetrability. My arguments will turn on the property of inaccessibility.[8]

As Carruthers ([2006], pp. 5–ff.) points out, we must be very careful not to conflate informational encapsulation with inaccessibility. Although both functional properties impose restrictions on the information flow between systems, they are not equivalent, as the following definitions make clear:

> Encapsulation: If a module M_1 is encapsulated in relation to another module M_2, its internal operations cannot draw on any information held in M_2.
>
> Inaccessibility: If a module M_1 is inaccessible to another module M_2, the latter has no access to the internal processing in M_1, but only to its outputs.

A component is not encapsulated if it can query other systems 'for information relevant to solving the task at hand.' (Carruthers [2006], p. 11) Yet, this system might be inaccessible to other systems, as would occur if they only had access to its outputs.

Cognitive impenetrability only lends support to encapsulation claims. When presented with the Müller-Lyer lines, visual modules can calculate the relative length of the lines without drawing on information from cognitive systems. This explains why the belief that the two lines are of the same length cannot change the misleading appearance of the two lines. Yet, this is compatible with the existence of some accessibility relations from cognitive into perceptual systems.

Interestingly, Pylyshyn not only holds that early vision is encapsulated from cognitive systems. He also thinks that it is inaccessible to them. As far as I can see, he endorses this stronger claim for two reasons: because he follows the common tendency of depicting encapsulation and inaccessibility as two sides of the same coin and, more importantly, because he responds to the previous transcendental arguments by stipulating that the referent of indices is fixed in a direct, non-cognitive, and causal manner. And this is sufficient to make indexing mechanisms inaccessible to cognition. Let me elaborate.

Most characterizations of encapsulation and inaccessibility present them as distinct properties (Fodor [1983], [2000]; Prinz [2006]; Robbins [2010]). Nevertheless, the way encapsulation is usually explained predicts that, if a module M_1 is encapsulated in relation to another module M_2, M_1 is inaccessible to M_2. No encapsulation without inaccessibility. This follows from the flowcharts that many cognitive scientists use to depict the lack of information relations between encapsulated systems and the rest of cognition. Thus, my first point is that a widespread flowchart analysis of the perception–cognition interface makes it the case that, if a perceptual system is encapsulated in relation to the cognitive system, that perceptual system is inaccessible to it. For the sake of clarity, I shall call this type of flowchart analysis 'the relay race model of modularity.'

In a 4x100 meter relay race, members of a team take turns running until they reach the finishing line. Since each runner must hand off the baton to the next runner within a certain zone, each participant makes a non-overlapping contribution to the shared goal of winning the race. Similarly, many defenders of encapsulation think of modules as related to each other as a *relayer* and a *relayee*. The relayee module must 'wait' for the relayer before it makes its contribution to the task at hand. This happens when a module cannot compute the value of a function until other modules deliver the required argument. In these cases, the module that is 'waiting' for the argument does not have any access to the intermediary steps needed to calculate it, in the same way as the next runner waits for the baton within a certain zone. In this model, encapsulation and inaccessibility overlap: the relayer cannot benefit from the relayee's contribution when it is covering its own distance (i.e., it is encapsulated) because the former bears a non-overlapping relation to the latter (i.e., it is inaccessible).[9]

This model is implicit in most debates on the (im)penetrability of early vision by cognition, which originate from Fodor's ([1983], [2000]) and Pylyshyn's ([1984], [1999]) seminal discussion. Defenders of the claim that early vision is encapsulated have tried to explain away putative examples of cognitive penetration by making one of the following three moves:

(1) *locating* the cognitive effects at a *pre*perceptual level;

(2) *locating* the cognitive effects at the *post*perceptual level; or

(3) *redefining* the cognitive effects as *internal* to the perceptual system.[10]

Crucially, these options can only exhaust the space of options if one is tacitly relying on a model according to which *any* information relation that goes from a module M_2 into a module M_1 entails that M_1 is unencapsulated in relation to M_2. This model is also hinted at in the dominant ways of characterizing encapsulation and inaccessibility. ‘Encapsulation’ is often used to mean that a system’s proprietary information is *independent* from other systems (Lyons [2001], p. 281, p. 297). In many cases, it is also assumed that encapsulation *entails* inaccessibility. Thus, Prinz ([2006], § 8) thinks that encapsulation ‘implies that one system cannot be accessed by another.’ In a similar vein, Robbins ([2010], § 1) defines inaccessibility as ‘[t]he flip side of informational encapsulation.’[11]

Interestingly, the relay race model is not only widespread in discussions on cognitive (im)penetrability; it is also *entailed* by the use of visual indices to block the regress that would be generated if reference were always satisfactional (section 1.2). Recall Pylyshyn’s contention that one can block the regress if one posits the existence of indices whose referent is fixed in a direct, non-cognitive, and causal manner. This hypothesis blocks the existence of any accessibility relations from the cognitive system to the indexing mechanism. Since the mechanism that fixes the referent of indices lies beyond the cognitive level, there is no room for informational relations from cognitive into indexing systems.

Let me clarify this point in light of Pylyshyn’s ([2007], pp. 68–9) use of interrupts to illustrate how the referent of visual indices is causally fixed. In computer science, an interrupt is a high priority signal sent to a processor that tells it to stop what it is doing, and attend to the interrupt. Interrupts can be generated in a number of ways. A good example is pressing a key. When one presses a key, a signal tells the computer that a keyboard event has occurred but does not tell it which key was pressed. Pylyshyn uses this analogy to explain how the referent of indices could be fixed in a causal and preconceptual manner. The cognitive system is related to indices in the same way as a processor is related to a keyboard. The former is wired to the latter *only* to *respond* to its high priority signals. If an interrupt is sent from the keyboard, an interrupt controller assigns it an order of priority, and then the cognitive system reads the interrupt. This analogy is a version of the relay race model of modularity. After all, it depicts the processor as a 'relayee' of the interrupt controller, so it can neither modify the workings of the keyboard, nor access the event that triggered the interrupt. All in all, if indexing mechanisms work like interrupts, they must be inaccessible to the cognitive system.

To sum up, Pylyshyn's hypotheses yield a two-tiered account. First of all, there is a direct, non-cognitive, causal mechanism responsible for *individuating* objects. Once they are individuated, they can be *identified* or *recognized*. In identification, the indexed item is paired with a symbol stored in long-term memory. In other words, perceptually based thoughts about *ordinary* objects result from the binding of indexed visual objects to conceptual predicates.

There are some reasons to be unsatisfied with this proposal. It is implausible to connect concepts *directly* with a single modality (vision), and ignore the role of cross-modality in our understanding of ordinary objects (Bullot [2009]). Nevertheless, we do not need to worry about this issue here. In what follows, I show that, even if defenders of the present conception of indices had something to tell us about cross-modal interactions, they would still get the perception–cognition interface wrong.

3. The Interface Problem

Since object files are prior to object recognition, they do not provide the *whole story* about the ability to refer to *ordinary* objects (section 1.1.). Thus, any model of mental architecture has to explain how acquaintance with entities satisfying these object constraints offers scaffolding for thoughts about ordinary objects. This is what I call the 'interface problem.' In what follows, I start with an intuitive characterization of this problem, and then present it in a more general way.

Suppose you posit a representation of changes in light intensity as a step in the representation of bounded objects. In order to provide a complete account of object perception, you will have to explain the sorts of transformations required to move from an input in terms of changes in light intensity to an output in terms of bounded objects. A solution to this problem can be provided from different perspectives. From a processing point of view, you have to sketch a flowchart analysis specifying how the transformations occur. From a conceptual point of view, you have to specify the conditions under which a representation *as of* changes in light intensity can ground a representation *as of* a bounded object. Whereas the former task is usually taken to be a topic of cognitive science, the latter is studied under the heading of the 'metaphysics of content.'

A similar problem arises in the present case. If you posit indices to connect mind and world, you have to explain how indices connect up with symbols in a language of thought (LOT). Imogen Dickie formulates this problem:

> [A]nyone who wants to build an account of acquaintance-based thought around empirical results about attention must deal with a gap between the class of potential objects of attention […] and the class of objects that most of our thoughts about the external world concern (ordinary material things). (Dickie [2010], pp. 221–2)

Dickie's problem is a version of a more general problem for any model of mental architecture that posits multiple processing levels. The issue is to make clear how the various levels connect up with the others. If a specific flowchart analysis of the perception–cognition interface prohibits some relations we have good reasons to posit, this is a good reason to reject it. Similarly, if the conditions to move from one level to the next are deemed insufficient, we have good reasons to reformulate the account accordingly.

I will argue that Pylyshyn's modularist interpretation of visual indices is not in a position to solve the interface problem. To this end, I shall focus on two aspects thereof: First, I will contend that the conception of indices as having their referent fixed in a bottom-up manner is incompatible with some forms of top-down attention. In other words, his flowchart analysis of the perception–cognition interface is incompatible with the existence of instructional effects on selective attention. Second, I will argue that indices do not provide sufficient conditions to ground thoughts about ordinary objects.

3.1. Top-Down Attention and Modularity

I explained in section 2 why the hypothesis of indices leads to a conception of modules as encapsulated and inaccessible to cognitive systems. Their encapsulation follows from the common idea that, if a mechanism *M* cannot be overturned by the cognitive states of a system, it is encapsulated in relation to the cognitive system. Their inaccessibility follows from the dominant way of explaining encapsulation in flowchart analyses of the perception–cognition interface and, more importantly, from the way visual indices block the regress arguments. According to Pylyshyn, and other authors, there is no regress in the way visual systems refer to the world because the visual system exploits preconceptual indices whose referent is fixed in a causal manner. Crucially, this blocks any accessibility relation from cognitive systems into indexing mechanisms.

This picture is inadequate, though. Consider top-down forms of attention. Top-down attention is exercised when the current *goals* or *cognitive states* of the system *modulate* the selection of information (Desimone and Duncan [1995]; Vecera [2000]; Pylyshyn [2003], p. 186; [2007], p. 56; Campbell [2002], p. 17; Clark [2006], p. 174). This kind of attention is a precondition of most experiments on visual tracking. In a typical MOT experiment, subjects are instructed to focus their attention on the flashed targets. Since subjects *understand* the instructions, some of their concepts get activated. Their conceptual understanding must then interact with the operation of indices if subjects are to follow the instructions. Nevertheless, this interaction couldn't take place if the cognitive system had no access to indexing mechanisms, or so I shall argue.

An initial difficulty arises if one depicts top-down attention as a voluntary action. We think of voluntary actions as involving a choice between alternatives, and make our choices from our knowledge of *what* alternatives are presented to us. Still, this conception does not fit the indexing view. As Pylyshyn ([2007], p. 95) observes, if voluntary attention involved knowledge of what the alternatives are, one would be led to conceive of the options as falling under concepts. This path is not open here, however, for indices operate in a causal, preconceptual manner.

The problems do not stop here, though. The relay race model of encapsulation is incompatible with the kind of top-down attention involved in typical experiments on visual reference. Consider a case in which a subject is asked to locate the two dots within the same closed contour (as in Figure 1, a). Since she understands the verbal instructions, the visual routine is conceptually initiated. This understanding will be causally efficacious only if the conceptual system can bear a top-down informational relation to the mechanisms responsible for scanning and indexing the dots. But, how can this conceptual process guide the searching process, if the conceptual system only has access to the output of indexing modules?

Interestingly, there is empirical evidence for the existence of top-down neural pathways in vision (Felleman and Van Essen [1991]; Desimone and Duncan [1995]; Vecera [2000]; Ruff [2011]). Pylyshyn himself grants that these pathways might be correlated with attention. Nevertheless, given his commitment to the relay race model of modularity, he is led to the implausible conclusion that attention operates *prior* to early vision:

> Even if such effects ultimately originate from 'higher' centers, they constitute a form of influence […] prior to the operation of early vision—that is, they constitute an early attentional selection of relevant properties. (Pylyshyn [2003], p. 68; see also: p. 89, [1999], p. 345, p. 358, [2003], p. 160, [2007], p. 44)

It is hard to make sense of this analysis, as the schema depicted in Figure 4 demonstrates. Indices can be characterized as functions that take a signal as input and return an argument as output. If they are like transducers, the signal must be described in a physical vocabulary. If not, their input must be the symbol delivered by transducers.

[Insert Figure 4]

Consider first the hypothesis that indices are functionally different from transducers. If their mechanism is inaccessible to cognition, there are only three ways of understanding top-down attention. If the conceptual system operates on the output of transducers (a), it can select some items without the intervention of visual indices. In other words, there is a form of visual selection that does not rely on indices, undermining the indexing hypothesis. If it operates on the values of the indexing mechanism (b), it will arrive *too late*. After all, these values are delivered *after* indexing occurred. Another option would be to claim that conceptual systems *already* operate on the worldly objects (c). But this view would be utterly implausible, for it would make no use of visual indices. Moreover, if we assume that the selected input is an object, we are led to a Gibsonian form of direct realism, a picture that allows the system to directly 'pick up' objects with no intervening computational processes. But defenders of Fodorian modularity typically reject this view (Fodor and Pylyshyn [1981]).

The situation would not improve if, *per impossibile*, one identified indices with transducers. If there is a cognitive interaction between symbolic states and physical input, cognition is not confined to the realm of representations. It extends, so to speak, to the world. To be sure, some theorists have advocated similar views (Clark [2008]). Nevertheless, these conceptions have been promoted as an *alternative* to more traditional accounts of cognition.

As a result, the only remaining option is to claim that cognitive states (not necessarily beliefs but goals or preferences) influence an *intermediary mechanism* of indexing. The trouble here is that this move would violate the putative *inaccessibility* of indexing mechanisms to cognition. This conclusion follows from the two available interpretations of this proposal. First, if there is a *direct influence* on the *inter-levels of processing responsible for visual indexing,* inaccessibility is violated. After all, this means that cognitive systems do not only have access to the outputs of indexing but also to the intermediary steps that generate those outputs. Second, suppose that the cognitive system can influence the indexing mechanism indirectly, i.e. by influencing a further mechanism that affects the inter-levels of processing responsible for visual indexing (as suggested by Macpherson [2012] in her treatment of the cognitive penetration of color experience). This indirect mechanism would violate inaccessibility as well. After all, information transmission is transitive: If *A* transmits a piece of information *I* to *B*, and *B* transmits *I* to *C*, *A* transmits *I* to *C*. Consider an analogy. I can invite you to my wedding by telling you that you are invited to my wedding. Alternatively, I can ask Paul to tell you that you are invited to my wedding. Although the latter communication chain is more indirect than the former, there is no informational difference between them. From an information-theoretic point of view, if you go to my wedding, it is because my message reached you, and enabled you to act accordingly. The informational relation between you and me holds independently of whether I used a direct or an indirect channel to transmit it (see Wu

[2013]).[12] But, as indicated above, both the relay race model of modularity and the hypothesis of indices prohibit *any* top-down informational relation from the cognitive system into the indexing mechanism.[13]

3.2. Selective Attention and Information

Another aspect of the interface problem is to explain how perceptually based thoughts about ordinary objects are grounded in primitive relations to visual objects. This section seeks to show that there is no plausible way of solving this problem if one grants the Fodorian interpretation of visual indices. My argumentative strategy is the following: I criticize a family of approaches that have been formulated within the mental-file framework, and then show that the problems of these views generalize to other modularist models of indexing mechanisms.

Dickie ([2010], [2011]) uses findings on visual tracking to explain conceptual reference. Though it is not clear whether she endorses the modularist framework, she accepts the relational model of perception, and takes it as the starting point to build a novel account of acquaintance with ordinary objects. Besides, she proposes to bridge the gap between visual and ordinary objects within a theory of mental files that is very similar to the one presented in section 2. Each object file is governed by object constraints. So my file for the complex demonstrative ‘this apple’ is governed by some principles that refer to visual objects. Still, a problem arises for this view: How does the cognitive system move from acquaintance with this apple qua visual object to acquaintance with this apple qua apple? More precisely: What conditions must be satisfied for the causal, primitive individuation of visual objects to ground thoughts about ordinary objects?

A natural option would be to claim that *attention* bridges the gap. Consider Campbell's view. He thinks attention underlies the 'intentional interrogation' of the environment (Campbell [2002], p. 3, p. 27, pp. 33–4, p. 89). On his view, attention is what makes available the 'categorical thing itself.' ([2002], p. 10, pp. 137–45, pp. 250–4) Unfortunately, Campbell gives us no hints on how attention performs this feat.

Consider now Dickie's proposal. She agrees with Campbell that attention may be invoked to bridge the gap between acquaintance with visual objects and thoughts about ordinary objects. On her view, however, one must impose some conditions on the etiology of the information relation that holds between the ordinary object and the cognitive system:

> [I]f a file of beliefs I form by taking what is delivered through an attentional channel at face value is about an object at all, it is about the object I am attending to—the object which is the source of the information in the file. […] So attending to a visual object puts you in a position to have acquaintance-based thoughts iff the visual object is an ordinary object. (Dickie [2010], p. 234)

If my perceptual state is *about* a visual object, and it happens to be *caused* by an ordinary object, I am in a position to have thoughts *about* an ordinary object. But how can attention perform this feat? If attention can bridge the gap between visual and ordinary objects, it must be able to select some *high-level properties* or 'pieces of information.' In other words, by directing my attention to the apple in front of me, I should be able to highlight, not only a continuous and persistent body, but also an apple. On Dickie's view, this is possible because the apple is the causal source of the visual object I am acquainted with.

In what follows, I argue that any use of attention-cum-information to bridge the gap between visual and ordinary objects betrays a misunderstanding on the semantic notion of information. Given that there is no other way of bridging this gap within the modularist interpretation of visual indices, the argument shows that the latter does not yield an adequate grounding of thoughts about ordinary objects.[14]

The problem here parallels Fodor and Pylyshyn's ([1981], pp. 186–ff.) objections to Gibson's theory of a direct pickup of information. As Fodor and Pylyshyn rightly point out, information is *relational*. A state of affairs S_1 contains information about another state of affairs S_2 provided their properties are non-accidentally correlated. If the tree contains information about its age, the latter must co-vary with another property, such as the number of rings. Similarly, when one says: 'those spots mean measles,' the properties of two events are non-accidentally correlated: the property of HAVING SPOTS and the property of HAVING THE MEASLES. Crucially, a cognitive system can only exploit these pieces information if it has a prior means of referring to the states of affairs or objects that are non-accidentally related. As we shall see, *this prevents the notion of information from bridging the gap between visual and ordinary objects*. Let me elaborate.

On Campbell's view, attention makes the categorical thing itself available. If the relation is conceived in causal terms, however, his liberal view only makes sense if one construes the notion of information as an intrinsic property of the entity one is related to. The information is directly available to the subject because it is 'revealed' through the perceptual contact with the thing.

Interestingly, the situation does not improve in Dickie's view. Recall that, on her view, in order to have perceptually based thoughts about the apple, it suffices to be perceptually acquainted with a visual object that is caused by an apple. The apparent plausibility of this move seems to originate from the ambiguity of 'about.' When I have experiences about a visual object, and the object happens to be an apple, my experiences are *of* this apple. The 'of' here indicates causal ancestry. But I cannot move from the weak claim that my perception is *of* this apple in the causal sense to the claim that I am in a position to form a belief *about* this apple. This would conflate the 'of' of causality with the 'about' of intentionality.

Consider an analogy. Suppose that Pierre calls you and says: 'Hi, it's me!' In this case, your auditory experience is *of* Pierre in the weak sense that it originates from Pierre. Moreover, your auditory experience *carries* information of Pierre. Still, if you do not recognize Pierre, you cannot form any perceptually based thought *about* Pierre. Being causally related to Pierre is not sufficient for you to be in a position to form a perceptually based thought about Pierre.

It would not do to refine the causal account by adding the further condition that, in order to be in a position to form thoughts about Pierre, one has to possess a concept of Pierre. Think of those cases in which you fail to recognize *familiar* people on the phone. You may try quite hard to focus your attention on the voice and, still, this will not help you know *who* is talking at the far end of the causal chain. Your experience is certainly about an auditory object whose causal source happens to be Pierre. Nevertheless, this is not sufficient for your cognitive system to transform the auditory signal *of* Pierre into a perceptually based thought *about* Pierre.

The voice in the phone carries information *of* Pierre. Nevertheless, its carrying information *of* Pierre does not 'reveal' that *Pierre* is speaking. On the relational view of information, it is not sufficient that an event (e.g., the voice in the phone) carries information of another event (e.g., Pierre's act of speaking) to reveal the information carried (Pierre's act of speaking). What is missing? First, one needs an independent grasp of the two terms of the information relation. Second, one has to be able to move from the presentation of one term of the relation to the other. Any correct theory of identification should elucidate the ability to make those moves.

It might be asked why these objections are taken to undermine the modularist account of visual indices. After all, I only considered two representative options. The reply is that these views display a *structural problem* in the modularist account of visual indices, so the prior objections generalize. Theorists of visual indices are trying to build a theory of object recognition *on the sole basis* of a primitive, causal relation to visual objects. Since relations to visual objects are the only 'windows' of the system to outer objects, it is a mystery how it can move from visual acquaintance with visual objects to thoughts about ordinary objects.

This argument generalizes to any formulation of visual indices within a view of indexing mechanisms as inaccessible to cognition. On this model, symbols are anchored to the world by means of causal relations to *visual* objects. Visual objects offer too *thin* a basis to ground thoughts about ordinary objects, for ordinary objects stand in a many-to-one relation to visual objects. Interestingly, none of the two resources available within modularist accounts of perception allow us to bridge the gap. First, if the processes are *barely causal*, as the transcendental arguments suggest, one has to construe information as an *intrinsic* property of events, i.e. as a property that can be 'read off' from the direct confrontation with one of the two terms of the relation. This mistake is implicit in Gibson's theory of a direct pickup of information, and we find it again in the magic account of attention as an ability to focus on high-level properties or pieces of information. Second, if the process is *inferential*, as Fodor and Pylyshyn ([1981]) contend, it is hard to see how the system can *form* and *verify* its hypotheses on ordinary objects on the sole basis of its acquaintance with visual objects. Suppose we want to explain how it forms the perceptually based thought: THAT IS A MOUNTAIN. In order to avoid trivialization, there must be limits to the properties that can be directly detected by transduction. Thus, the property of being a mountain cannot be directly 'given.' This creates a problem: How does the system verify the hypothesis that *that* is a mountain, if it only has access to low-level properties and visual objects? If the transition is inferential, it will lack the means of verifying its high-level hypotheses.[15]

I am not aware of any other solution to the previous objections within the modular account of perceptual systems. If defenders of these views think there is a better alternative, I am all ears.

4. Revising the Indexing Hypothesis

I offered two arguments against the introduction of indices whose referent is fixed in a purely causal way. Since the interface problem has no solution within this framework, we should question the transcendental arguments. Those arguments purport to show that, *if* cognitive processes are transformations of symbols, there must be deictic anchors to the world. One can resist the arguments in two ways: by rejecting the premise (i.e., some cognitive processes operate on predicates) or revising the conclusion (i.e., the hypothesis of visual indices). I will pursue the latter strategy.

The transcendental arguments clearly show that not all reference can be fixed by the satisfaction of properties, so some form of singular reference is necessary to relate perception and cognition to the world. Similarly, the empirical findings show that human visual systems have distinctive *tracking* abilities. Yet, these findings do not lend support to Pylyshyn's stronger claim that the reference of indices is fixed in a causal manner, i.e. in a way that lies beyond the purview of a theory of cognition. Pylyshyn ([2001]) is probably right when he writes: 'Sooner or later concepts must be grounded in a primitive causal connection between thoughts and things.' Yet, it is not clear that the mechanisms that select and keep track of objects are *solely* grounded in a causal connection. Although it is a requirement that transducers be *data driven*, as Pylyshyn ([2007], p. 42) stresses, this need not be the case of the early vision system!

If all we need to ground concepts in perception is *some* causal constraints, there is nothing to prevent us from imposing those causal constraints only at the preobjective level of transduction. This would leave room for accessibility relations between cognitive and indexing systems. In what follows, I make some suggestions on how the flowchart analysis of the perception–cognition interface could be modified along these lines.

4.1. Revising the Perception–Cognition Interface

In order to ground conceptual awareness of ordinary objects in perceptual relations to visual objects, some defenders of indices liberalize the notion of attention, and construe information as an intrinsic property of objects. Since this picture is to be avoided, we should ask why it seems so attractive to a number of authors. My first hypothesis is that this picture becomes unavoidable when one conceives of cognitive processes as transformations of symbols *independent* from perceptual modules.

Recall the transcendental arguments. They assume that cognitive processes operate on symbols, and proceed to bridge the gap between these symbols and the world. In this case, one has to assume that cognition is independent from perception. This assumption becomes apparent in Pylyshyn's contention that the referent of indices is fixed by primitive causal relations to the world. I think we can make some progress if we redefine the relation of cognition to perception.

Consider an analogy. Some vending machines are designed to detect various coins. Nevertheless, those machines do not perform this feat by detecting high-level properties like BEING A DIME or BEING A QUARTER. Instead, their designers exploit some reliable correlations between these properties and low-level properties that machines do detect, such as shape, size, and width. This explains why one can defraud them by introducing other coins that exemplify the same low-level properties. Thus, these machines detect coins *by detecting* low-level properties. What allows a machine to make the transition from the detection of low-level properties to the detection of high-level properties is its *design*. Vending machines are *designed* to treat things that satisfy a number of low-level properties as exemplifying other high-level properties.

Human perceptual systems bear some similarities to vending machines. Even though they lack a designer, we can pursue the analogy, and treat natural selection and experience as producing functional changes. Perceptual systems were wired by natural selection to recognize ordinary objects by detecting visual objects. They can make this transition because they are wired to exploit some reliable correlations of low-level object constraints with high-level properties of ordinary objects. Yet, they differ from vending machines in a crucial respect: they can *broaden* their recognition abilities in the course of experience. What humans have that vending machines lack is plasticity, i.e. the capacity to change their functional organization through experience.

This remark suggests a way out. The hypothesis that cognition needs causal anchors to the world leads us to overlook the role of plasticity. It is as if concepts were already there, 'waiting' to be causally linked to the world. This creates a mystery as to how these concepts can be 'hooked' to ordinary objects. This problem should not arise, however, if we had a picture of experience that reshapes the perception–cognition relation. When the conditions under which the functional modification took place are fulfilled, the transition from relations to visual objects to awareness of ordinary objects should be automatic. When these conditions are not fulfilled, however, the transition should not take place. This would happen if you are used to identifying Pierre by his way of walking and his face, but are presented only with the voice in the phone.[16]

4.2. Revising the Modularity of Early Vision

The second assumption that prevents the modularist view of indexing mechanisms from handling the interface problem is its conception of modular visual systems as inaccessible to cognition. This is clear in the picture of indexing as a purely bottom-up process. This model makes it difficult to see how conceptually driven forms of attention are possible. My second suggestion is to explore new accessibility relations between perceptual and cognitive modules that are compatible with facts of cognitive impenetrability.

What the psychophysical tests for cognitive impenetrability show is that some perceptual mechanisms are cognitively rigid, for they cannot be easily modified by background beliefs. Nevertheless, there is no direct path from cognitive inflexibility to the claim that the underlying mechanism is connected with subsequent mechanisms by delivering the arguments of the function they compute. Actually, that conceptually articulated instructions can have an impact on selective attention strongly suggests that there must be *top-down* content-sensitive interactions from beliefs into indexing mechanisms. The mistake is to think that, if a module M_1 has access to another module M_2, M_1 must be capable of modifying M_2.

One can illustrate this idea by comparing the cognitive system to an enterprise. The manager could entertain two sorts of relations with the subordinates: each should report to her about the tasks it has done and—on the basis of these reports—the manager could give them new assignments. The manager might be unable to change *what* the subordinates do. She might only know how important their job is for the goals of the company but lack the know-how to change their operations. Nevertheless, her knowledge could be used to guide the employees by imposing deadlines or reorganizing their schedules. In this case, she might modify *some ways* in which the subordinates do their job, even though it could not directly affect *what* they do.

This point is well illustrated by some competitive conceptions of attention. On these accounts, top-down processes can influence the operations of low-level brain regions.

A number of theorists have described these forms of attention as producing a ‘target template’ that *facilitates* or *biases* low-level perceptual processing (Duncan and Humphreys [1989]; Bundesen [1990]; Desimone and Duncan [1995]; Vecera [2000]; Dosher and Lu [2010]; Ruff [2011]). These models typically hypothesize that a cognitive signal is sent to low-level processing areas. This signal might be compared to a description that becomes active in working memory. This description confers an advantage to some targets over the others. The description enables the system to *enhance* or *amplify* the firing rate of some cells and *inhibit* or *attenuate* that of others. As a result, there is a contrast-gain amongst stimuli. If the system is faced with a cluttered scene, some populations of neurons will be disposed to respond to the target items as though the strength of their stimulation had increased, while others will be disposed to display an attenuated response to distractors, as if their stimulation had attenuated.[17]

Let me translate these suggestions into the present model. Suppose a subject is asked to focus on the two apples in a cluttered scene. Her conceptual understanding leads her to form the intention: FOCUS ON THE TWO APPLES. Since the intention is too sophisticated, it cannot directly guide the visual search. After all, object files lack ‘understanding’ of apples. Nevertheless, the intention might be at the origin of a *partial* communication chain with lower-level mechanisms. This might occur if those systems could decode only part of the message. One option would be that they ‘understand’ the concept TWO, as when a foreign speaker grasps a single word in a sentence of another language. In this case, the activation of the concept TWO might be sufficient to trigger the activation of two potential object files. Though the original message carries more information than lower systems can ‘take,’ what those systems can take might be sufficient for them to become active. Once the two potential object files become active, the system would be led to increase their activity in those areas responsible for detecting continuous and persistent bodies. When the modules that feed object files individuate two continuous and persistent bodies, the activation pattern of object files would change, and send a new signal to the cognitive systems. In some cases, the firing pattern would be similar to those cases where the system formed thoughts about apples. Thus, the concept APPLE would be activated as well. In other cases, however, the firing pattern would differ from that of successful apple recognition, leading the system, either to activate a different concept (e.g., ORANGE in a case of misrecognition), or send a new query to the perceptual modules (e.g., KEEP SEARCHING).[18]

I do not claim that this simple system describes the actual working of top-down attention. Nevertheless, it illustrates how the relevant ideas from competition models of attention might be exploited to revise the indexing hypothesis. First, it shows that, in order to make room for top-down forms of attention, one has to give up the crude conceptual–preconceptual and cognitive–causal divides. If cognitive states modulate activity in perceptual systems, there must be accessibility relations between them. Perceptual systems can be guided by conceptual understanding because the latter may produce signals that are sufficient to generate neural activity that matches the cognitive goals of the system. I illustrated this point by describing a communication chain in which low-level systems partially ‘understand’ high-level signals. This solution is not open to defenders of Fodorian modularity because they think of perceptual systems as inaccessible to cognition. And it is not open to the present version of visual indices because they operate at a preconceptual level. Second, the present proposal shows that concepts are not detached symbols that receive the arguments delivered by visual indices. Instead, it strongly suggests that conceptual systems lie in a continuum with perceptual systems.

5. Concluding Remarks

Humans have distinctive mechanisms of visual tracking. What is controversial is the assumption that those mechanisms are similar to indices that refer to objects in a purely causal manner. I showed that this view faces two problems: it cannot explain how conceptually initiated forms of attention can guide visual scanning and its separation of concepts from perception makes it hard to explain how awareness of visual objects works as scaffolding for thoughts about ordinary objects.

I also suggested that we do not need to think of reference to objects as a purely bottom-up process. In section 4.1, I proposed to conceive of experience, not only as a causal input to cognition, but also as modifying the relation of perception to cognition. In section 4.2, I hypothesized that partial communication chains from cognition to perception could make room for instructional effects on selective attention. The two solutions suggest an overall picture of perceptual experience that does not move in a single direction, from the world to cognition, but operates in cycles from perception to cognition, and vice versa. Thus, one can avoid the regress problems, not by positing a causal leap to the world, but a non-vicious circularity that enables cognition to guide perceptual reference and allows experience to create new links between perception and cognition. If this view is on the right track, the distinctions between bottom-up and top-down are not clear-cut; they are a matter of emphasis or perspective.[19]

This picture casts doubt on the starting point of the transcendental arguments: a conception of cognitive processes as *isolated* from perception and the world. We should rather start our inquiry by considering the mind as *situated* in the world, gaining new concepts through its perceptual interaction with the environment and adapting its relation to the world to its own goals. If we managed to explain the emergence of concepts from this mind–world interaction, the need of connecting them through causal relations should not arise. If our starting point were a situated mind, the urgent task would not be to bridge the gap but to explain how that gap emerges.

Funding

Swiss National Science Foundation (research grant No. 100015_131794).

Acknowledgments

Thanks to Jérôme Dokic and Clare Mac Cumhaill, who read previous versions of this material and made some insightful comments, to Ariel Cecchi for conversations on attention and cognitive penetration, and to two anonymous referees for their challenging criticism. Figures 1–3 are reproduced with permission of the publishers.

Département de philosophie

Université de Genève

Santiago.Echeverri@unige.ch

References

Bach, K. [1987]: *Thought and Reference*, Oxford: Clarendon Press.

Ballard, D. H., Hayhoe, M. M., Pook, P. K. and Rao, R. P. N. [1997]: 'Deictic Codes for the Embodiment of Cognition', *Behavioral and Brain Sciences*, **20** (4), pp. 723–67.

Bermúdez, J. L. [2003]: *Thinking Without Words*, New York: Oxford University Press.

Biederman, I. [1995]: 'Visual Object Recognition', in S. M. Kosslyn and D. N. Osherson (*eds*), *An Invitation to Cognitive Science*, 2nd edition, vol. 2: Visual Cognition, Cambridge, MA: MIT Press, pp. 121–65.

Borg, E. [2004]: *Minimal Semantics*, Oxford: Clarendon Press.

Bullot, N. [2009]: 'Toward a Theory of the Empirical Tracking of Individuals: Cognitive Flexibility and the Functions of Attention in Integrated Tracking', *Philosophical Psychology*, **22** (3), pp. 353–87.

Bundesen, C. [1990]: 'A Theory of Visual Attention', *Psychological Review*, **97**, pp. 523–47.

Campbell, J. [2002]: *Reference and Consciousness*, Oxford: Clarendon Press.

Carey, S. [2009]: *The Origin of Concepts*, New York: Oxford University Press.

Carruthers, P. [2006]: *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought,* Oxford: Clarendon Press.

Cecchi, A. [forthcoming]: 'Cognitive Penetration, Perceptual Learning, and Neural Plasticity', *dialectica.*

Churchland, P. M. [1988]: 'Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor', *Philosophy of Science*, **55**, pp. 167–87.

Clark, A. [2008]: *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford / New York: Oxford University Press.

Clark, A. [2000]: *A Theory of Sentience*, New York: Oxford University Press.

Clark, A. [2006]: 'Attention and Inscrutability: A Commentary on John Campbell, "Reference and Consciousness" for the Pacific APA Meeting', *Philosophical Studies*, **127** (2), pp. 167–93.

Dennett, D. C. [1991]: *Consciousness Explained*, London: Penguin Books.

Deroy, O. [forthcoming]: 'Modularity of Perception', in M. Matthen (*ed.*), *The Oxford Handbook of Philosophy of Perception*, New York: Oxford University Press.

Desimone, R. and Duncan, J. [1995]: 'Neural Mechanisms of Selective Visual Attention', *Neural Review Neuroscience*, **18**: 193–222.

Dickie, I. [2010]: 'We Are Acquainted with Ordinary Things', in R. Jeshion (*ed.*), *New Essays on Singular Thought*, New York: Oxford University Press, pp. 213–45.

Dickie, I. [2011]: 'Visual Attention Fixes Demonstrative Reference by Eliminating Referential Luck', in C. Mole, D. Smithies and W. Wu (*eds*), *Attention: Philosophical and Psychological Essays*, New York: Oxford University Press, pp. 292–321.

Dosher, B. and Lu, Y.-L. [2010]: 'Attention: Theories of', in E. B. Goldstein (*ed.*), *Encyclopedia of Perception*, vols. 1 and 2, Los Angeles, CA: Sage Publications, pp. 107–12.

Dretske, F. [1981]: *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.

Driver, J. and Frith, C. [2000]: 'Shifting Baselines in Attention Research', *Nature Reviews Neuroscience*, **1**: 147–8.

Duncan, J. and Humphreys, G. W. [1989]: 'Visual Search and Stimulus Similarity', *Psychological Review*, **96**: 433–58.

Felleman, D. J. and Van Essen, D. C. [1991]: 'Distributed Hierarchical Processing in the Primate Cerebral Cortex', *Cerebral Cortex*, **1** (1): 1–47.

Fodor, J. A. [1983]: *The Modularity of Mind*, Cambridge, MA: MIT Press.

Fodor, J. A. [1984]: 'Observation Reconsidered', in his *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, pp. 231–51.

Fodor, J. A. [2000]: *The Mind Doesn't Work that Way*, Cambridge, MA: MIT Press.

Fodor, J. A. [2008]: *LOT 2: The Language of Thought Revisited*, New York: Oxford University Press.

Fodor, J. and Pylyshyn, Z. [1981]: 'How Direct Is Visual Perception? Some Reflections on Gibson's "Ecological Approach"', *Cognition,* **9**, pp. 139–96.

Intriligator, J. and Cavanagh, P. [2001]: 'The Spatial Resolution of Attention', *Cognitive Psychology*, **4** (3), pp. 171–216.

Jackson, F. [1977]: *Perception*, Cambridge: Cambridge University Press.

Kahneman, D., Treisman, A. and Gibbs, B. [1992]: 'The Reviewing of Object Files: Object-Specific Integration of Information', *Cognitive Psychology*, **24**, pp. 175–219.

Lyons, J. [2001]: 'Carving the Mind at its (Not Necessarily Modular) Joints', *The British Journal for the Philosophy of Science*, **52**, pp. 277–302.

Lyons, J. [2011]: 'Circularity, Reliability, and the Cognitive Penetrability of Perception', *Philosophical Issues*, **21**, pp. 289–311.

Macpherson, F. [2012]: 'Cognitive Penetration of Color Experience: Rethinking the Issue in Light of an Indirect Mechanism', *Philosophy and Phenomenological Research,* **84** (1), pp. 24–62.

Mole, C. [2009]: ‘Attention’, in E. N. Zalta (*ed.*), *Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*, <http://plato.stanford.edu/archives/fall2009/entries/attention/>

Olivers, C. N. J., Watson, D. G. and Humphreys, G. W. [1999]: ‘Visual Marking of Locations and Feature Maps: Evidence from Within-Dimension Defined Conjunctions’, *Quarterly Journal of Experimental Psychology A,* **52A** (3), pp. 679–715.

Peterson, M. A. [1994]: ‘Object Recognition Processes Can and Do Operate Before Figure-Ground Organization’, *Current Directions in Psychological Science*, **3**, pp. 105–11.

Prinz, J. [2006]: ‘Is the Mind Really Modular?’, in R. Stainton (*ed.*), *Contemporary Debates in Cognitive Science*, London: Wiley-Blackwell, pp. 22–36.

Pylyshyn, Z. W. [1984]: *Computation and Cognition: Toward a Foundation of Cognitive Science*, Cambridge, MA: MIT Press.

Pylyshyn, Z. W. [1999]: ‘Is Vision Continuous with Cognition? The Case for Cognitive Impenetrability of Visual Perception’, *Behavioral and Brain Sciences*, **22** (3), pp. 341–423.

Pylyshyn, Z. W. [2000]: ‘Situating Perception in the Visual World’, *Trends in Cognitive Sciences*, **4** (5), pp. 197–207.

Pylyshyn, Z. W. [2001]: ‘Visual Indexes, Preconceptual Objects, and Situated Vision’, *Cognition*, **80** (1/2), pp. 127–58.

Pylyshyn, Z. W. [2003]: *Seeing and Visualizing: It’s Not What You Think*, Cambridge, MA: MIT Press.

Pylyshyn, Z. W. [2007]: *Things and Places: How the Mind Connects with the World*, Cambridge, MA: MIT Press.

Raftopoulos, A. [2001]: ‘Is Perception Informationally Encapsulated? The Issue of the Theory-Ladenness of Perception’, *Cognitive Science*, **25**, pp. 423–51.

Raftopoulos, A. [2009]: *Cognition and Perception: How Do Psychology and Neural Science Inform Philosophy?*, Cambridge, MA: MIT Press.

Recanati, F. [2012]: *Mental Files*, New York: Oxford University Press.

Robbins, P. [2010]: ‘Modularity of Mind’, in E. N. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition),* <http://plato.stanford.edu/archives/sum2010/entries/ modularity-mind/>

Ruff, C. C. [2011]: ‘A Systems-Neuroscience View of Attention’, in C. Mole, D. Smithies and W. Wu (*eds*), *Attention: Philosophical and Psychological Essays*, New York: Oxford University Press, pp. 1–23.

Samuels, R. [2006]: ‘Is the Human Mind Massively Modular?’, in R. Stainton (*ed.*), *Contemporary Debates in Cognitive Science*, London: Wiley-Blackwell, pp. 37–56.

Scholl, B. [2007]: ‘Object Persistence in Philosophy and Psychology’, *Mind and Language*, **22** (5), pp. 563–91.

Scholl, B. [2009]: ‘What Have We Learned about Attention from Multiple-Object Tracking (and Vice Versa)?’, in D. Dedrick and L. Trick (*eds*), *Computation, Cognition, and Pylyshyn*, Cambridge, MA: MIT Press, pp. 49–77.

Scholl, B. and Pylyshyn, Z. W. [1999]: ‘Tracking Multiple Items through Occlusion: Clues to Visual Objecthood’, *Cognitive Psychology*, **38**, pp. 259–90.

Scholl, B., Pylyshyn, Z. W. and Feldman, J. [2001]: ‘What is a Visual Object: Evidence from Target-Merging in Multiple-Object Tracking’, *Cognition*, **80**, pp. 159–77.

Shea, N. [forthcoming]: ‘Distinguishing Top-Down from Bottom-Up Effects’, in S. Biggs,

M. Matthen and D. Stokes (*eds*) *Perception and Its Modalities*, Oxford: Oxford University Press.

Spelke, E. S. [1990]: ‘Principles of Object Perception’, *Cognitive Science*, **14**, pp. 29–56.

Stern, R. [2011]: ‘Transcendental Arguments’, in E. N. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy (Fall 2012 Edition)*, <http://plato.stanford.edu/archives/fall2012/entries/transcendental-arguments/>

Stokes, D. [2013]: ‘Cognitive Penetrability of Perception’, *Philosophy Compass*, **8/7**, pp. 646–63

Treisman, A. [1998]: ‘Feature Binding, Attention and Object Perception’, *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, **353** (1373), pp. 1295–306.

Treisman, A. and Gelade, G. [1980]: ‘A Feature Integration Theory of Attention’, *Cognitive Psychology*, **14**, pp. 107–41.

Ullman, S. [1984]: ‘Visual Routines’, *Cognition*, **18**, pp. 97–159.

Vecera, S. P. [2000]: ‘Toward a Biased Competition Account of Object-Based Segregation and Attention’, *Brain and Mind*, **1**, pp. 353–84.

Xu, F. [2003]: ‘The Development of Object Individuation in Infancy’, in H. Hayne and J. Fagen (*eds*), *Progress in Infancy Research*, vol. 3, Mahwah, NJ: Erlbaum, pp. 159–92.

Wu, W. [2013]: ‘Visual Spatial Constancy and Modularity: Does Intention Penetrate Vision?’, *Philosophical Studies*, **165**: 647–69.

APPENDIX

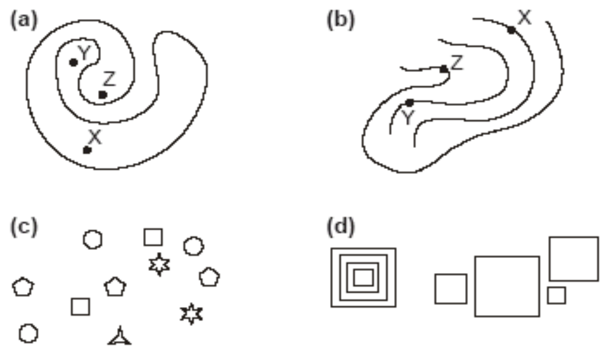

Figure 1 Examples of tasks requiring visual routines: (a) What points are inside a closed contour? (b) What points are in the same curved line? (c) Which objects are collinear? (d) How many squares are there in each group? Figure borrowed from Pylyshyn (2000).

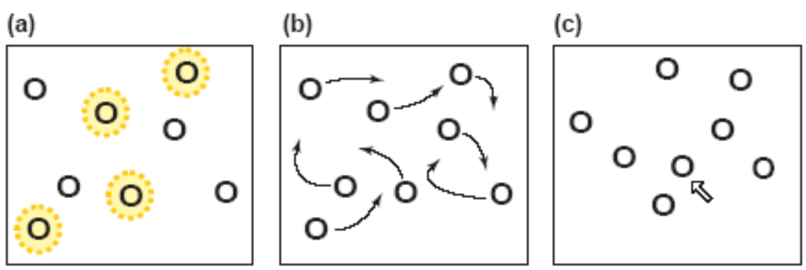

Figure 2 Schematic representation of a typical MOT experiment: (a) eight circles appear on the screen and four of them are flashed to indicate their status as targets; (b) next, circles move randomly for 10 seconds; (c) at the end, subjects are asked to select the four targets with the mouse. Figure borrowed from Pylyshyn (2000).

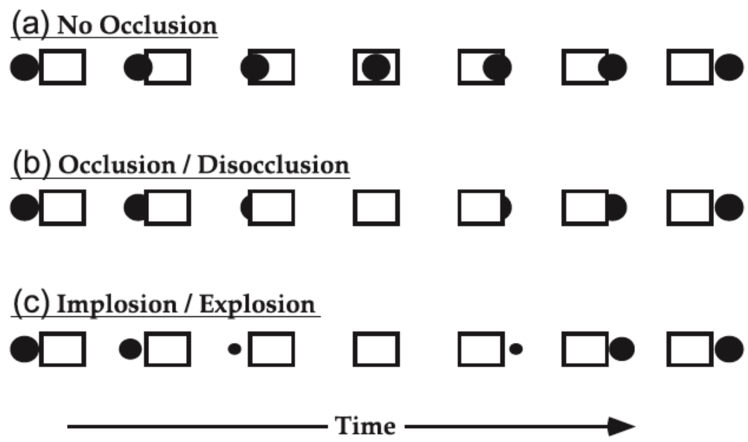

Figure 3 Representation of the ways objects might interact with an occluder (Scholl & Pylyshyn 1999). (a) The object does not disappear but moves atop the other contours. (b) The object is momentarily occluded, deleting from its trailing edge as it appears and then accreting from its leading edge upon its reappearance. Tracking in this condition is robust. (c) The object moves exactly as in (b), but it gradually 'implodes' to a point, and then 'explodes' from a point. Tracking is greatly impaired due to spatiotemporal discontinuity. Figure borrowed from Scholl (2007: 568).

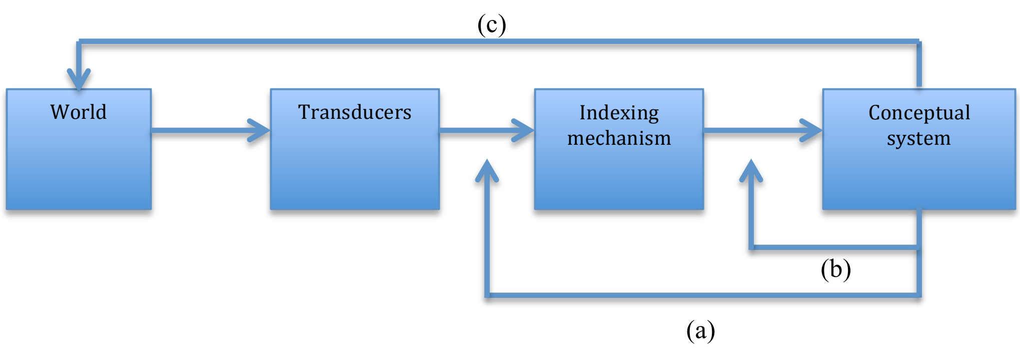

Figure 4 Making sense of the conceptual guidance of indices: (a) the conceptual system intervenes on the output of transducers, (b) on the output of the indexing mechanism, (c) on the world.

ENDOTES

[1] The qualification is meant to exclude obvious counterexamples, as when one says (pointing to a map): 'This is my hometown.'

[2] Pylyshyn ([2003], [2007]) does not consider the possibility that non-satisfactional reference might be realized in complex demonstratives, such as 'That colored object.' This nicety does not matter for our present purposes, provided we grant that an indexical element is needed to block the regress. He also furnishes some processing arguments to show that visual reference cannot be secured by description but I do not have the space here to examine them.

[3] This assumption might be questioned, though. Some theorists think that a feature can be encoded and still be inaccessible to conscious report. Such views are based on the assumption that the brain may represent properties 'in multiple, parallel and overlapping systems, only some of which put the subject in a position to think about the properties they represent.' (Mole [2009], § 2.3) A generalization of this idea is Dennett's ([1991]) Multiple Draft Model of Consciousness.

[4] Studies on prelinguistic infants and non-human animals confirm these findings (Spelke [1990]; Carey [2009]). The emerging consensus is that vision scientists and developmental psychologists have been studying the same *psychological kind*. Nevertheless, whereas developmental psychologists describe the underlying mechanisms as part of a conceptual system, Pylyshyn describes them as part of early vision. Although it is not clear how to characterize the output of the early vision system, Pylyshyn ([1999], p. 361) relies on the widespread idea that it includes a representation of visible surfaces similar to Marr's 2½ D

sketch. See also Raftopoulos ([2009], p. 51) and Cecchi ([forthcoming]). This interpretation seems to be confirmed by the experience of occlusion in tracking experiments. This is all I need for my present purposes.

[5] This includes modularists who hold that not all modules are Fodorian but still think that some of them are encapsulated and/or inaccessible (Carruthers [2006]; Deroy [forthcoming]).

[6] I will not consider other features of Fodorian modules.

[7] On the distinction between individuation and identification, see Biederman ([1995]), Kahneman *et al.* ([1992]), and Xu ([2003]). Although most theorists assume that individuation is a precondition of identification, this view is not universally accepted. An example of dissent is Peterson ([1994]), who thinks that familiarity effects on figure–ground segregation suggest that object recognition may be prior to object individuation.

[8] A number of theorists use the adjectives 'cognitively impenetrable' and 'encapsulated' as synonyms (see Lyons [2011]). This use is incorrect, however, for a perceptual system can be *cognitively* impenetrable but penetrable to other *perceptual* systems. Besides, we need to distinguish the phenomenon from its explanation. I am using 'cognitively impenetrable' to describe the fact that the operation of some principles cannot be (easily) overturned by *cognitive* states. The adjective 'encapsulated' explains this fact by imposing some restrictions on the information relations that obtain between impenetrable and cognitive systems. This need not be the only possible explanation, though. If one rejects the notion of a module—as Prinz ([2006]) and Samuels ([2006]) do—one must explain cognitive impenetrability in a different way.

[9] As far as I know, the relay race metaphor has not been used in the literature to describe this type of flowchart analysis of the perception–cognition interface. Still, there is strong evidence that this model is pervasive. Deroy ([forthcoming], § 3a) uses a structurally similar analogy to describe the encapsulation of early vision. On her view, if a module M_1 is encapsulated in relation to another module M_2, M_1 is like a 'two-way hermetic tube': 'Nothing escapes before the output, and nothing gets in except at the level of inputs.' Clearly, this model predicts that, if M_1 is encapsulated in relation to another module M_2, M_2 cannot have access to M_1. Deroy's analogy gets support from the way Fodor ([1983], pp. 41–2, p. 56, p. 60, p. 69) himself presents encapsulation. And, although Fodor ([2000], pp. 113–4) acknowledges that there are different ways of implementing encapsulation, he thinks of them as notational variants, sticking to a definition of encapsulation that takes the relay race model for granted ([2000], p. 63). Similarly, in his discussion of Fodor's notion of encapsulation, Lyons [(2001, p. 296] writes that '*s* is informationally encapsulated *from r* iff any communication between *s* and *r* is strictly unidirectional. […] This notion of informational encapsulation forbids any top-down information exchange.' The same relay race model shapes Borg's ([2004], p. 133, pp. 142–ff.) modular account of semantic understanding. See also Raftopoulos ([2001]).

Pylyshyn ([2007], p. 56) compares the operation of indexing mechanisms to 'the way you might pick out a fish by placing a baited hoot in the water—it happens primarily at the initiative of objects; we say it is data driven.' As will become clear later, this analogy is equivalent to a relay race model of indexing mechanisms. Carruthers ([2006], pp. 219–20) approvingly presents an 'enzymatic model' of the perception–cognition interface that, if taken literally, is structurally similar to the relay race model as well.

[10] See Fodor ([1983]), Churchland ([1988]), Pylyshyn ([1999], [2003]) and various references therein, Raftopoulos ([2001], [2009]), Stokes ([2013], pp. 654–ff.), and Deroy ([forthcoming]). Lyons ([2011], pp. 303–ff.) distinguishes five foci of possible influence of cognition on vision. Yet, none of these options offers a different alternative from the ones considered here. His option A corresponds to my option 1, his option B describes cases of genuine penetration, and his options C, D, and E are variants of my option 2.

[11] Why is the relay race model so pervasive in analyses of informational encapsulation? It is probably because it is the simplest way of explaining the encapsulation of a system in relation to another system. If you want to explain why a system cannot rely on information held outside of it to produce its output, depict it as informationally independent or isolated from other systems. Another reason is that many people seem to think that unavailability to verbal report is a reliable indicator of inaccessibility to cognition. To my knowledge, this assumption has not been adequately defended.

[12] Philosophers working on cognitive penetration often point out that cognitive influences on perception must be direct in order to be non-trivial. This requirement is meant to exclude cases in which a cognitive state causes a set of actions that then cause an experience, as when one's desire to watch the tennis game motivates one to turn on the TV, thereby occasioning a visual experience (see Stokes [2013], p. 647). The sort of indirect mechanism examined here is not trivial in this sense. Although it is not an example of informational

unencapsulation, it satisfies the usual requirement that cognitive states bear a semantic relation to visual states (Pylyshyn [1999], p. 343; Macpherson [2012], p. 26).

[13] One might wonder why the defender of the modularist interpretation could not *just* say that top-down effects of attention on perceptual processes are compatible with the encapsulation of early vision. My reply is that the devil is in the details. Sure, some defenders of Fodorian modularity have recognized that top-down effects of attention are compatible with impenetrability (Pylyshyn [1999]; Raftopoulos [2009], pp. 303–4). In order to block the present objection, however, one needs to provide an alternative flowchart analysis of encapsulation that makes room for the relevant top-down influences and keeps inaccessibility. Unfortunately, this model is not available either to those who think of encapsulation within a relay race model, or to those who block the regress of transcendental arguments by conceiving of reference fixation as a bottom-up causal matter. That is why they insist that top-down effects either (1) influence the input to early vision, (2) or the output of early vision, (3) or are internal to the early vision system. For discussion, see Fodor ([1983]), Pylyshyn ([1999]), Raftopoulos ([2001], [2009]), and Deroy ([forthcoming]). If my arguments are correct, instructional effects on selective attention require accessibility relations of cognition into indexing systems.

[14] The phrase 'semantic notion of information' is used to denote the family of analyses of information that originate from Dretske's *Knowledge and the Flow of Information.* Although inspired by Shannon's mathematical theory, these approaches differ from Shannon's because they offer an account of *what* information is transmitted. Besides, these notions are taken to be semantic because the sentences used to describe the information

carried by a signal are inten*s*ional (with an 's'). For a recent use of this semantic notion of information in the literature on cognitive penetration, see Wu ([2013], p. 649, p. 656, p. 662).

[15] Fodor and Pylyshyn ([1981]) purported to show, *contra* Gibson, that because there is no direct pickup of information, perception is inferential. The present considerations suggest that this conclusion is not warranted.

[16] Churchland ([1988]) also emphasizes the role of plasticity, and complains that defenders of encapsulation tend to take the mental architecture as fixed. This charge clearly applies to Pylyshyn. Yet, my suggestion is only to introduce plasticity to explain how perceptual awareness of visual objects can ground conceptual awareness of ordinary objects. The suggestion is to conceive of experience, not only as a causal 'leap' to visual objects, but also as a source of functional changes in the perception–cognition interface.

[17] These authors report modulation of firing rates at early stages of visual processing within the primary visual cortex (V1). The areas associated with top-down control of attention seem to be located in the amygdala and the posterior parietal and frontal cortices.

[18] My picture of accessibility as *partial communication* is consistent with Raftopoulos' ([2001], [2009]) notion of semi-encapsulation, which he uses to denote off-line top-down influences of cognition on perception. My view differs from his in a crucial respect, though: it is based on a flowchart analysis of the notions of encapsulation and inaccessibility, so it treats them as independent properties. In my view, it is somewhat misleading to treat off-line accessibility as an attenuated form of encapsulation.

Notice also that the present view provides some elements to rebut Carruthers' ([2006], p. 59) contention that most modules are inaccessible to each other. His argument is based on the claim that, if a module M_1 is accessible to another module M_2, the latter must contain a model of M_1 or be capable of simulating it. But this contention lacks support, for accessibility only requires *partial* communication. Although I grant Carruthers' distinction between encapsulation and inaccessibility, I reject his contention that accessibility is a *rare* phenomenon. Actually, accessibility seems necessary to account for many top-down forms of control.

[19] Shea ([forthcoming]) argues for a similar view but on the basis of different arguments.