The Role of Visual Representations in Seeing

Michael Barkasi*

APA Central 2019†

Abstract

Many cognitive psychologists and some philosophers hold that the construction of a representation of a thing within the visual system is necessary for, or constitutive of, seeing it. I argue against this view: you can have a visual experience of something without a representation of that thing being constructed within your visual system. But I don't go so far as to argue that the visual system isn't in the business of constructing representations or that visual experiences themselves aren't representational. To argue for my claim I give a specific example of something seen, but for which there's no representation constructed in the visual system. The example is certain parts of objects used in a MOT experiment from Brian Scholl et al. The results of the experiment show that no visual representation is constructed; I then argue, from our ability to voluntarily attend to those parts, that they are visually experienced.

1 Introduction

A tempting view is that seeing constitutively—and so necessarily—involves your visual system constructing a representation of what's seen (e.g., Tye 1995: 100–23; Prinz 2000: 249, 2006: 454, 2011: 174; Wagemans et al. 2012: 1180). In this paper I argue that this view is false. Specifically, I give an example of a thing you see but for which no representation is constructed in the visual system. The example is seeing certain arbitrary parts of the figures used in a particular multiple object

*Postdoctoral Research Fellow at the Network for Sensory Research, University of Toronto Mississauga. michael.barkasi@utoronto.ca

†For presentation at the upcoming 2019 APA Central meeting, colloquium paper. Denver. February 20–23.

tracking (MOT) task from a study by Brian Scholl, Zenon Pylyshyn, and Jacob Feldman (2001). I show that they're seen but not represented in three steps.

Step 1: Using the results from Scholl et al., I argue that during the MOT task the visual system does not construct a representation of certain parts of the tracked figures.

Step 2: Using introspection, I argue that during the MOT task these nonrepresented parts are available for voluntary attention.

Step 3: By appealing to a standard view of voluntary attention, I argue that because they are available for voluntary attention these nonrepresented parts are seen.

If the argument works it's plausible that the visual system does not construct representations of other arbitrary parts of seen objects, in other cases, but these parts are seen. The thesis here is specifically about seeing in which the perceiver has a visual experience of what's seen. So the thesis is that the construction of a representation of a thing by the visual system is not constitutive of having a visual experience of it.

Note that what's at issue isn't whether seeing constitutively involves the construction of representations by the visual system. I take for granted that seeing does involve these (as I'll sometimes say for short) visual representations.[1] Instead, what's at issue is the relationship between visual representations and visual experience. If successful, the argument here shows there isn't a one-to-one mapping between things seen and visual representations. Instead, the visual system engages in a fairly limited construction of representations which underlies seeing a rich array of things.

Also note that the question posed here is orthogonal to the debate between representationalism and naïve realism. Denying that seeing a thing constitutively involves visual representations of it is consistent with representationalism. Few representationalists assume that having a visual experience of a thing constitutively involves a visual representation of it (Michael Tye and Jesse Prinz are exceptions, see above). In addition, both the assumption that consciously seeing a thing constitutively involves some visual representation and the further (here rejected) claim that it involves a visual representation of the thing experienced are consistent with denying that experiences are representations and accepting that they're

[1] This assumption is rejected by ecological psychologists, enactivists, and some working in AI (e.g., Gibson 1966, 1986; Brooks 1991; van Gelder 1995; Noë 2004; Hutto and Myin 2013; see also Akins 1996; Orlandi 2011a,b, 2014).

relations (e.g., Campbell 2002: 119). The nature of experiences—representational or relational—need not be grounded directly in what happens in the visual system.

Zenon Pylyshyn has also argued that a visual representation of a thing is not constitutive of having a visual experience of it (2007: 120–23; see also Chalmers 2000; Noë 2004; Noë and Thompson 2004; Dennett 1978; McDowell 1994; and Mitroff et al. 2005). His argument starts with his FINST account of MOT results. As he presents it, the FINST account entails that at any one time only a few objects are represented by the visual system (those being tracked by a FINST). But, he suggests, introspectively it seems that you have a panoramic visual experience of many objects. So, there must be some disconnect between what's represented in the visual system and what's consciously seen.

My argument makes several advances. First, although step 1 also draws on MOT work, I don't assume Pylyshyn's FINST-based account of those results. Second, I use the empirical results to point to a specific example of something not visually represented: specific parts of a figure during a particular MOT task. Third, my approach to arguing that these parts are visually experienced does not rest on the folk panorama view of visual experience. Steps 2 and 3 of my argument show that these parts are visually experienced without appealing to any introspective judgments about visual experience.

2 Step 1: Visual System Representations

2.1 Preliminaries on Visual Representations

What is the visual system? Cognitive psychology gives an explanation of how organisms see by breaking it into simpler, subpersonal subtasks (Drayson 2012). Each of these subtasks is functionally defined and carried out by neural processes in the brain. Often these subtasks are to compute some output as the function of some input. The *visual system* is the collection of neural processes realizing these functionally defined subtasks. It is *constructive* in sense that the functions it computes take as inputs (e.g.) encodings of retinal stimulation and have as outputs encodings of the distal causal sources of that stimulation—of what's seen. What neural activity in the retina directly encodes is the distribution of light intensity on the retina. Seeing requires working backwards from encoded distributions of light intensity to their causes, i.e. to what's seen (e.g., Marr 1980: 203; see also Fodor and Pylyshyn 1981). The input and output states of these computations are representational: the input and output encodings can be more or less accurate and are about what's seen (Burge 2010; Orlandi 2014). David Marr's work on shape

extraction provides a well known example of this constructive process (Marr and Nishihara 1978; Marr 1980; 1982).

The construction of representations by the visual system can be divided into two operations: feature extraction and grouping (Kahneman et al. 1992: 176–8; Scholl 2001: 16; Scholl et al. 2001: 160; Wagemans et al. 2012: 1180; see also Clark 2004). Feature extraction involves the detection of a feature (e.g., shape, color, or orientation) based on input from an earlier stage of processing (or directly from retinal stimulation). Marr's work, for example, provides an account of shape extraction. Grouping involves *binding* clusters of those features together so as to treat them as belonging to the same thing (see figure 1). Grouping results in a *segmentation* of features in the scene into discrete chunks, or into distinct perceived things. On an influential account (Treisman and Gelade 1980; Kahneman et al. 1992; Treisman 1998), this grouping process is temporally extended and involves tracking a changing cluster of features over time in an "object file". The distinction between grouping and feature extraction is relative to a level of processing (Wagemans et al. 2012: 1188). For example, shape extraction, as described by Marr (1980: 211), itself involves grouping of previously extracted features (grouping edges when moving from the primal sketch to the 2½D sketch, and grouping shape parts when moving to the 3D model).

The point is that the construction of a representation of a thing in the visual system happens either when it's a feature that's extracted by the visual system, or is a thing the features of which the visual system groups together. The output states of feature extraction and grouping operations represent (respectively) the extracted features and the things which have the grouped features. If you want to know whether the visual system constructs a representation of something seen, the question is whether that thing is an extracted feature or has features the visual system groups together.

2.2 MOT and Arbitrary Object Parts

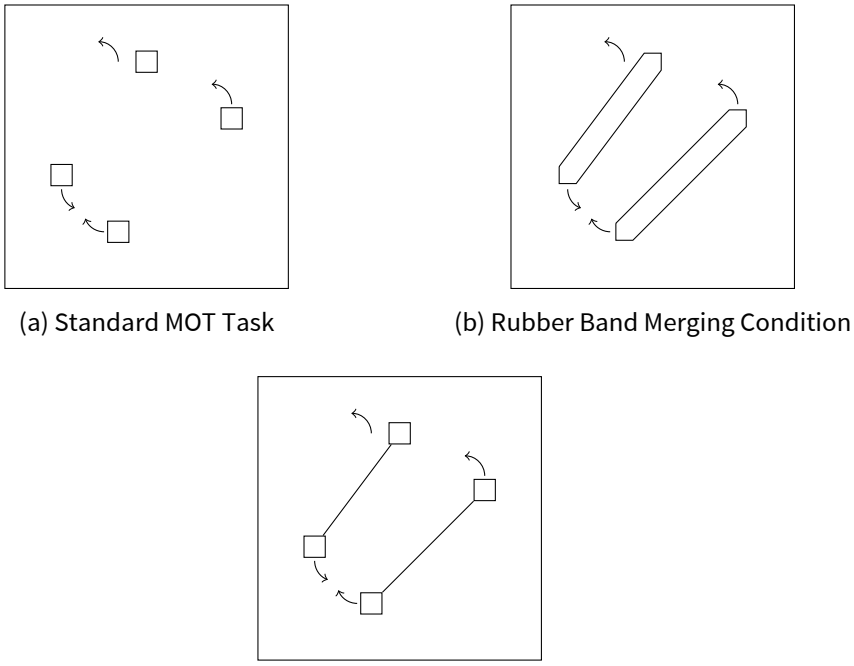
The MOT paradigm (Pylyshyn and Storm 1988; Pylyshyn 2007; Scholl 2001) provides behaviorial, nonintrospective tests for whether the features of a seen thing are being grouped together, and so represented, by the visual system (Scholl et al. 2001: 161). In MOT tasks a subject looks at a computer screen which displays a number of "objects" (around eight), e.g. crosses, disks, or squares, none of which have any features that would allow them to be distinguished from the rest. Some of the objects (about four), called *targets*, flash or in some other way are cued. Then all the objects, including the uncued distractors, move in random, independent

Figure 1: Gestalt Grouping Demonstration: Dalmatian Photo. Michael Bach reports on his website that the image is from Gregory (1970) (photo by Ronald James), but first published in *Life Magazine*, 2/19/1965, p. 120. See http://www.michaelbach.de/ot/cog_dalmatian/.

paths within the screen. When the objects stop the subject must pick out the target objects, for example by clicking on them with the mouse (Pylyshyn 2007: 34–35). The interesting result is that subjects can do this task at all, and with relative ease (Scholl 2009: 59). Further, performance remains relatively constant when there is up to four or five target objects, and then drops off significantly after that.

What's crucial for the discussion here is that the ability to track the targets indicates that the features of those targets (edges, shape, location, color) are being grouped together, and hence that the visual system is constructing a representations of the targets. So, a failure to successfully complete a given MOT task indicates that the visual system is not constructing representations of the targets (Scholl 2001: 32; Scholl et al. 2001: 171–72). This is suggested by the two features just mentioned: tracking in MOT tasks is relatively easy and performance is constant to a point after which it falls flat. Both are characteristic of a fast, automatic perceptual operation like grouping (Dickie 2010: 220).

A MOT experiment from Scholl, Pylyshyn, and Feldman (2001) provides the example of a thing not represented by the visual system. In their experiment they



(a) Standard MOT Task

(b) Rubber Band Merging Condition

(c) Dumbbell Merging Condition

Figure 2: Schematic Diagrams of Scholl et al. MOT Task. (a) shows a standard MOT set up: several distinct figures (usually around 8, 4 displayed here) move at random around the screen. Some of these are cued at the start as targets (usually around 4, we can imagine the bottom two were cued here) and the subject tracks them as they move. (b) shows the rubber band merging condition. In this task the ends (both targets and distractors) move exactly how they did in (a), but are "merged" to form the elongated rectangles. (c) shows the dumbbell merging condition. Figures based on/adapted from (Scholl et al. 2001: fig.1 and fig.2).

took the usual eight objects in a MOT task (eight boxes), paired each of the four targets with a distractor, then "merged" the paired targets and distractors (see figure 2). They used eight different merging conditions.[2] One of the merging conditions, the one which provides the example, wraps the boxes in a solid line, as if they've been wrapped in a rubber band (figure 2b). Another, mentioned here as a contrasting case, joins the boxes with a solid line so that the merged boxes form a dumbbell shape (figure 2c). The key result is that on some merging conditions, e.g. the rubber band condition, subjects are unable to track the targets, while on others,

[2] See http://www.yale.edu/perception/Brian/demos/MOT-Merging.html.

e.g. the dumbbells, tracking remains possible (Scholl et al. 2001: 170–72, fig.3). So subjects watch the now merged boxes move randomly around the screen, and in some merging conditions are able to select the boxes (the ends of the merged pairs) that were initially cued as targets. In other merging conditions (e.g., the rubber band condition) subjects cannot reliably select at the end which boxes (which ends of the merged pairs) were initially cued. As noted above, the inability to track in the rubber band merging condition suggests that the merged target and distractor boxes are no longer being grouped separately as distinct objects. Instead, the visual system treats the merged target-distractor box pair as a single object: it groups together all the features of the target box, distractor box, and rubber band around them as a single object. Call these merged target-distractor pairs *TD pairs*.

So, the target box-ends in the rubber band merging condition provide an example of things for which no visual representation is constructed (see figure 3, which displays the TD pairs and their ends). Specifically, no visual representation is constructed while you do the MOT task involving the TD pairs. Since feature grouping is task dependent, the failure to track the target box-ends only shows that visual representations of those target ends aren't constructed during the MOT task. It might be suggested that during the MOT task representations of the target ends are constructed, but just not usable in the task itself. Against this suggestion, there are no known grouping principles which would support the construction of representations of the target ends in the rubber band merging case (Scholl 2001; Wagemans et al. 2012), while there's empirical evidence against the construction of object-part representations in these cases (Poljac et al. 2012).

3 Step 2: Voluntary Attention

In this section I argue that the target ends of TD pairs are available for voluntarily attention. I claim that when you look at an object, it seems to you that you can voluntarily attend to any part of it which you can identify as a part. To support this claim about the possibility of attention to identified parts, consider again the TD pairs. When I look at the TD pairs in figure 2 (or watch the video of them moving) it seems plausible that I'm able to pick one TD pair and voluntarily attend to its target end. At least, I can do this once the target end is identified to me as a potential target (e.g., as it is on the right in figure 3). This introspective claim can be further supported by noting that it's very easy to track the target end of a single TD pair, a task which seems to require voluntarily attending to that target end. Note that the failure to successfully track all four target ends in this MOT task does not show that you cannot voluntary attend to the target ends. At best, failure

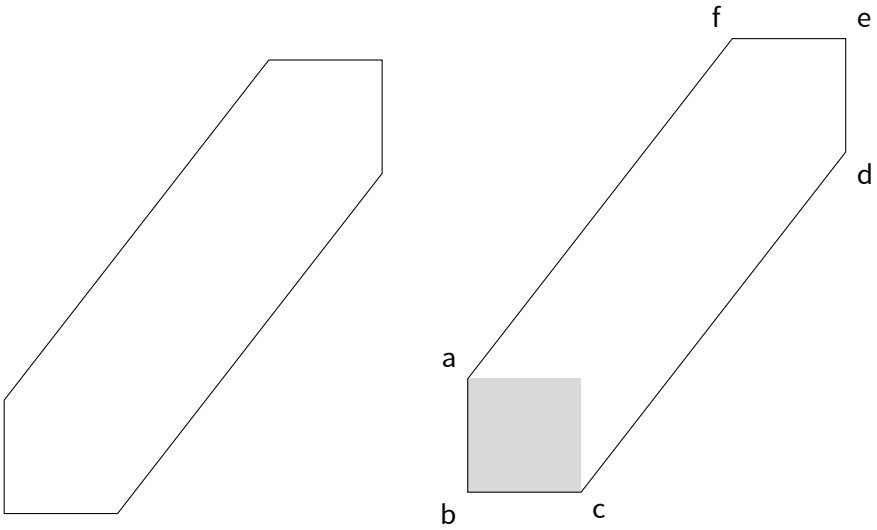


Figure 3: A single TD pair (left). The figure on the right labels the nodes of the TD pair and shades just the surface of the original target box. In constructing a representation of a TD pair the visual system groups together line segments ab, bc, cd, de, ef, and fa, along with the surface bounded between them. The target end of the TD pair to the left is the part corresponding to the shaded region and its adjacent line segments (lines ab and bc). The claim that the target ends are not represented in the visual system is the claim that these elements in the left TD pair (lines ab, bc, and the surface corresponding to shaded region on the right) are not grouped together when you look at it.

to track all four target ends in this task only shows that you cannot voluntarily keep attention directed to all four at once—it doesn't show that a single target end cannot be voluntarily attended.

How could you voluntarily attend to something not represented by the visual system? What you can voluntarily attend is not constrained by how your visual system segments a scene, i.e. by what is represented by the visual system. A high-level, conceptually mediated capacity for identification plus the intention-guided character of voluntary attention allows you to voluntarily attend to objects for which the visual system does not construct representations. Our visual system may not segment out some part of an object, but you have a post-visual cognitive capacity to do so. This explanation explains why tracking multiple target ends of TD pairs isn't possible, despite the possibility of tracking a single target end. If visual representations of the target ends were constructed, then attention to (and hence tracking of) them would would relatively unconstrained and effortless. But

attending to the target ends requires post-perceptual conceptual processing and this post-perceptual processing is relatively constrained (by, e.g., the limits of working memory, see Beck 2012; Hutchinson and Turk-Browne 2012). These constraints limit you to attending to only one of the target ends.

4 Step 3: From Attention to Visual Experience

This final section completes the overall argument by arguing that since the target ends of TD pairs are available for voluntary attention, you have visual experience of them. The key claim is that visual experience of a thing is necessary for it to be available for voluntary attention. This claim about the necessity of visual experience for the availability of voluntary attention follows from a widely held view about the relationship between attention and perception. The view, often tacitly assumed, is that voluntary attention is a mechanism which operates to select consciously perceived things (Valberg 1992: 21; Scholl 2001: 20; Pylyshyn 2007: 59; Levine 2010: 181; Dickie 2010: 216, 2011: 303 Wu 2011: 109; cf. Prinz 2011). On this view attention and conscious perception are different personal-level mental acts and their underlying subpersonal systems are functionally distinct. Objects and object parts are first perceptually experienced, and then attention is a distinct action which operates on those objects and their parts. If this view is correct, then what's available for voluntary attention are the things you consciously perceive. So in the case of vision, what's available are things you consciously see. So the availability of the TD pair target ends requires that they are consciously seen.

You might object that there's empirical evidence that visual experience is not required for the availability of voluntary attention. Robert Kentridge (2008; 2011) has done Posner-style cuing studies with the blindsight patient GY, which, he argues, show that GY attends to things for which he has no visual experience. If GY can attend to things he unconsciously sees, then experience of them was not required for their availability for attention. In these studies arrows are flashed in the intact portion of GY's visual field which point to areas in the blind field. Then a vertical or horizontal line is flashed in the blindfield and, as in standard blindsight tests, GY must guess the orientation of the line. Accuracy in the guessing task is better when location of the line is congruent with the direction of the arrow. The posited explanation is that the flashed arrow cues attention to the area in which it points, and this cued attention speeds reaction time. (Increased reaction time is a standard way to operationalize or measure attention in experiments.)

Kentridge's work does not pose a problem for the final step in the argument. First, there are alternative explanations of the results which don't posit that GY is

attending to areas of his blindfield. Prinz discusses several of these (2011: 193–94). But even if Kentridge is correct that there is attention involved, there's only a problem on a very strong and implausible interpretation. Specifically, there's only a problem if GY is voluntarily attending to the lines themselves. If the attention is involuntary or not directed at the lines themselves (and instead, e.g., is directed at spatial locations), then there's no problem. Both of these are plausibly true. First, like any other cuing task, the kind of attention involved in Kentridge's studies is presumably involuntary. It's not that GY sees the arrow and deliberately directs attention in the indicated direction. Instead, the cued arrow grabs attention.[3] Second, attention is being cued to spatial locations, not the lines themselves. Kentridge himself characterize the results this way, saying that "we had demonstrated selective spatial attention in blindsight" (2011: 239).

References

Akins, Kathleen. 1996. "Of Sensory Systems and the "Aboutness" of Mental States". *The Journal of Philosophy* 93: 337–72.

Beck, Jacob. 2012. "The Generality Constraint and the Structure of Thought". *Mind* 121: 563–600.

Brooks, Rodney. 1991. "Intelligence without representation". *Artificial Intelligence* 47: 139–59.

Burge, Tyler. 2010. *Origins of Objectivity*. Oxford University Press.

Campbell, John. 2002. *Reference and Consciousness*. Oxford University Press.

Chalmers, David J. 2000. "What is a Neural Correlate of Consciousness?" In Thomas Metzinger (ed.), *Neural Correlates of Consciousness*. MIT Press. Page citations to the reprint in David Chalmers. 2010. *The Character of Consciousness*. Oxford University Press.

Clark, Austen. 2004. "Feature-Placing and Proto-Objects". *Philosophical Psychology* 17: 443–469.

Dennett, Daniel. 1978. "Toward a cognitive theory of consciousness". *Minnesota Studies in the Philosophy of Science* 9: 201–28.

Dickie, Imogen. 2010. "We are Acquainted with Ordinary Things". In Robin Jeshion (ed.), *New Essays on Singular Thought*. Oxford University Press, 213–245.

——. 2011. "Visual Attention Fixes Demonstrative Reference by Eliminating Ref-

[3]GY claims he can direct attention to areas of his blindfield (Kentridge 2011: 239). But the setup in which the results were obtained is a cuing task.

erential Luck". In Christopher Mole, Declan Smithies, and Wayne Wu (eds.), *Attention: Philosophical & Psychological Essays*. Oxford University Press, 292–322.

Drayson, Zoe. 2012. "The Uses and Abuses of the Personal/Subpersonal Distinction". *Philosophical Perspectives* 26: 1–18.

Fodor, Jerry A. and Pylyshyn, Zenon W. 1981. "How Direct Is Visual Perception?: Some Reflections on Gibson's 'Ecological Approach'". *Cognition* 9: 139–196.

Gibson, James J. 1966. *The Senses Considered as a Perceptual System*. Houghton Mifflin Company.

——. 1986. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum.

Gregory, Richard. 1970. *The Intelligent Eye*. Weidenfeld & Nicolson.

Hutchinson, J. Benjamin and Turk-Browne, Nicholas B. 2012. "Memory-Guided Attention: Control From Multiple Memory Systems". *Trends in Cognitive Sciences* 16: 576–579.

Hutto, Daniel D. and Myin, Erik. 2013. *Radicalizing Enactivism: Basic Minds without Content*. The MIT Press.

Kahneman, Daniel, Treisman, Anne, and Gibbs, Brian. 1992. "The Reviewing of Object Files: Object-Specific Integration of Information". *Cognitive Psychology* 24: 175–219.

Kentridge, Robert W. 2011. "Attention Without Awareness: A Brief Review". In Christopher Mole, Declan Smithies, and Wayne Wu (eds.), *Attention: Philosophical & Psychological Essays*. Oxford University Press, 228–246.

Kentridge, Robert W., Nijboer, Tanja C. W., and Heywood, Charles A. 2008. "Attended but unseen: Visual attention is not sufficient for visual awareness". *Neuropsychologia* 46: 864–69.

Levine, Joseph. 2010. "Demonstrative Thought". *Mind and Language* 25: 169–195.

Marr, David. 1980. "Visual Information Processing: The Structure and Creation of Visual Representations". *Phil Trans R. Soc London B* 290: 199–218.

——. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Marr, David and Nishihara, Herbert K. 1978. "Representation and recognition of the spatial organization of three-dimensional shapes". *Proc. R. Soc. Lond. B* 200: 269–294.

McDowell, John. 1994. "The Content of Perceptual Experience". *Philosopical Quarterly* 44: 190–205.

Mitroff, Stephen R., Scholl, Brian J., and Wynn, Karen. 2005. "The Relationship

Between Object Files and Conscious Perception". *Cognition* 96: 67–92.
Noë, Alva. 2004. *Action in Perception*. The MIT Press.
Noë, Alva and Thompson, Evan. 2004. "Are there Neural Correlates of Consciousness?" *Journal of Consciousness Studies* 11: 3–28.
Orlandi, Nicoletta. 2011a. "Embedded Seeing: Vision in the Natural World". *Noûs* Published first online, December 29.
——. 2011b. "The Innocent Eye: Seeing-as without Concepts". *American Philosophical Quarterly* 48: 17–31.
——. 2014. *The Innocent Eye: Why Vision Is Not a Cognitive Process*. Oxford University Press.
Poljac, Ervin, de Wit, Lee, and Wagemans, Johan. 2012. "Perceptual Wholes can Reduce the Consciouss Accessibility of their Parts". *Cognition* 123: 308–312.
Prinz, Jesse. 2000. "A Neurofunctional Theory of Visual Consciousness". *Consciousness and Cognition* 9: 243–259.
——. 2006. "Beyond Appearances: The Content of Sensation and Perception". In Tamar Szabo Gendler and John Hawthorne (eds.), *Perceptual Experience*. Oxford University Press, 434–460.
——. 2011. "Is Attention Necessary and Sufficient for Consciousness?" In Christopher Mole, Declan Smithies, and Wayne Wu (eds.), *Attention: Philosophical & Psychological Essays*. Oxford University Press, 174–203.
Pylyshyn, Zenon W. 2007. *Things and Places: How the Mind Connects with the World*. The MIT Press.
Pylyshyn, Zenon W. and Storm, R.W. 1988. "Tracking Multiple Independent Targets: Evidence for a Parallel Tracking Mechanism". *Spatial Vision* 3: 179–197.
Scholl, Brian J. 2001. "Objects and Attention: the State of the Art". *Cognition* 80: 1–46.
——. 2009. "What Have We Learned about Attention from Multiple-Object Tracking (and Vice Versa)?" In Don Dedrick and Lana Trick (eds.), *Computation, Cognition, and Pylyshyn*. The MIT Press, 49–78.
Scholl, Brian J., Pylyshyn, Zenon W., and Feldman, Jacob. 2001. "What is a Visual Object? Evidence from Target Merging in Multiple Object Tracking". *Cognition* 80: 159–177.
Treisman, Anne. 1998. "Feature Binding, Attention and Object Perception". *Phil Trans R. Soc London B* 353: 1295–1306.
Treisman, Anne and Gelade, Garry. 1980. "A Feature-Integration Theory of Attention". *Cognitive Psychology* 12: 97–136.
Tye, Michael. 1995. *Ten Problems of Consciousness: A Representational Theory*

of the Phenomenal Mind. The MIT Press.

Valberg, J.J. 1992. "The Puzzle of Experience". In Tim Crane (ed.), *The Contents of Experience: Essays on Perception*. Cambridge University Press, 18–47.

van Gelder, Tim. 1995. "What might cognition be if not computation?" *Journal of Philosophy* 92: 345–81.

Wagemans, Johan, Elder, James H., Kubovy, Michael, Palmer, Stephen E., Peterson, Mary A., Singh, Mansih, and von der Heydt, Rüdiger. 2012. "A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure-Ground Organization". *Psychological Bulletin* 138: 1172–1217.

Wu, Wayne. 2011. "Attention as Selection for Action". In Christopher Mole, Declan Smithies, and Wayne Wu (eds.), *Attention: Philosophical & Psychological Essays*. Oxford University Press, 97–116.