Vision, reanimated and reimagined

Shimon Edelman[*]

June 22, 2012

Abstract

The publication in 1982 of David Marr's *Vision* has delivered a singular boost and a course correction to the science of vision. Thirty years later, cognitive science is being transformed by the new ways of thinking about what it is that the brain computes, how it does that, and, most importantly, why cognition requires these computations and not others. This ongoing process still owes much of its impetus and direction to the sound methodology, engaging style, and unique voice of Marr's *Vision*.

> We find certain things about seeing puzzling, because we do not find the whole business of seeing puzzling enough.
> — Ludwig Wittgenstein
> (*Philosophical Investigations*, 1958, II, xi, p.212f)

1 Introduction

Vision is harder than it looks. As with other cognitive faculties, the immediacy and ease of seeing belie the difficulties faced by those trying to understand how it works. Instructors in introductory AI courses like to tell the story of a summer project assigned in 1966 by MIT's Marvin Minsky to a group of undergraduates: to hook up a camera to a computer and write some software that would implement "a significant part of a visual system."[1] In AI, the frontal attack on vision that began with such skirmishes quickly devolved into trench warfare. By mid-1970s, whatever progress that computer vision had to show for a decade of hard work tended to concentrate in toy domains or "blocks worlds"; not so much in natural scenes, where one is forced to contend with what Sloman (1983) called "the horrors of the real world."

[*] Dept. of Psychology, Cornell University, Ithaca, NY 14850, USA. Address correspondence to edelman@cornell.edu

[1] See MIT AI Memo #100 (July 7, 1966), titled *The Summer Vision Project* and signed by Seymour Papert. Many of the undergraduates involved in that project, including the coordinator, Gerald Jay Sussman, went on to make major contributions to computer science and artificial intelligence.

Meanwhile, the rate of progress in understanding the brain basis of vision was hardly any more satisfying. Although neuroscientists may consider themselves to be closer to the real world than computer vision researchers,[2] the typical laboratory setups used to study visual perception in the brain were (and many still are) hardly more ecologically valid than AI's block worlds. Worse, because the brain is so overwhelming complex, poking around in it without a clear idea of the function that is being studied is a sure recipe for missing the proverbial forest for the trees. Thus, the initial excitement with the finding by Hubel and Wiesel, in early 1960s, of presumed neural representations of contour elements (Hubel and Wiesel, 1959, 1968), which, moreover, looked like possible building blocks for coding objects and scenes, has led to thousands of new studies and a flood of data, but little understanding of general principles of vision.

The work of David Marr, whose entire graduate training and research career spanned a mere decade and a half and culminated in the posthumous publication of his book *Vision* (1982), served as a game-changer both on the machine and on the biological side of vision research. Marr's contribution to the field extends far beyond specific theories advanced by his earlier publications (collected with commentaries in a volume edited by Vaina, 1991). It also transcends, I believe, the unparalleled impact that his book had on generations of scientists. The present article, intended as an homage to that contribution, takes the form of a series of some very high-level questions, of the kind that Marr excelled in putting forward and that his work made safe for all of us to play with. The questions are accompanied by brief answers, many of them synthesized from previously published empirical studies and reviews, where one can also find ample references in support of the views expressed here.

2 What does it mean, to see?

Marr's approach to science combined a Wittgensteinian propensity to ask disruptive foundational questions with a finely honed ability to seek answers and formulate explanations on a variety of levels. The vertically integrated inquiry, spanning levels from the abstract-computational, focusing on questions of purpose and utility, down to the implementational, which deals with neurons or with computer architectures, is, of course, the core of Marr's methodological contribution to brain and cognitive sciences, developed jointly with Poggio (Marr and Poggio, 1977). While a complete explanation of a particular visual faculty requires that all the levels be addressed, at the beginning of a project it is especially important to formulate a clear computational-level hypothesis that alone can serve to steer the inquiry and bind together the strands of explanation generated by it. I can think of no better way of making the first step in this direction than asking again the question that opens Marr's *Vision*: "What does it mean, to see?"

The observation that Marr offers in response to this question — "The plain man's answer (and Aristotle's,

[2] A joke about "the computer engineer, who, when asked to describe how he would write a computer program to recognize a chicken, replied, 'first, assume a spherical chicken' " even made it into print (Reynolds and Desimone, 1999, p.19).

too) would be, to know what is where by looking" — seems entirely uncontroversial, and so does the analysis that follows, in which a distinction is made between the *process* of vision and the *representation* of the world that this process computes.[3] And yet, both the answer and the distinction can — indeed, must, if we are to take Marr's own methodological legacy seriously — be skeptically examined.

2.1 Problems with the intuitive answer

The highly intuitive, "plain man's answer" to the fundamental question of vision contains a couple of hidden assumptions that need to be brought out into the open. As I have already written about these matters at some length elsewhere (Edelman, 1995, 1999, 2002, 2009), my goal here is merely to call these assumptions into question, in the hope of stimulating further discussion.

The first assumption, which I shall term ontological (cf. Akins, 1996), amounts to the belief that there is a matter of fact regarding that which is "out there" in the scene, and, moreover, that the putative facts are statable in a form that is both unambiguous and inherent to the scene itself (i.e., independent of the viewer). In highly impoverished laboratory situations, it may be possible to describe each scene in the form of a list of objects that jointly account for all of it, while leaving nothing out ("what"), accompanied by complete information about their locations ("where"). For natural scenes of any kind, such a description is, however, largely out of the question, simply because discrete, namable visual objects, along with any boundaries between them, are a matter of interpretation (that is, are relative to the observer and the quirks of the observer's representation language), not of absolute truth (Edelman, 2009).

The breakdown of the ontological assumption suggests that it is not a good idea to equate the purpose of vision with the recovery of 3D object-centered, part-based structure of the world (as Marr and Nishihara (1978) have done). It also explains in part why a theory of image understanding based on that idea (Biederman, 1987) did not fare well either as an account of human vision (Edelman, 1999) or as a recipe for machine vision approaches (Dickinson et al., 1997). Tellingly, a less ambitious version of the ontological project, which called for the recovery not of object-centered 3D geometry but of "intrinsic images" (Barrow and Tenenbaum, 1978) did not in practice fare well either (Barrow and Tenenbaum, 1993).[4]

[3] The realization that vision, being an information processing task, is inherently computational is another key contribution of the Marr-Poggio research program, and is also a key point in the introductory chapter of *Vision*. Unfortunately, this point still eludes some people who really should know better; witness the remark made by Marcus du Sautoy, Simonyi professor for the public understanding of science and a professor of mathematics at the University of Oxford: "...when I'm doing mathematics my brain is doing so much more than just computation. It is working subconsciously, making intuitive leaps." (*The Observer* on March 31, 2012).

[4] An intrinsic image is a representation that makes explicit certain objective ("intrinsic") properties of the viewed scene, such as the 3D orientation of the surfaces that comprise each object. To appreciate how problematic this idea is, note that even just parsing a scene into sets of pixels that "belong together," as in the LabelMe project (Russell et al., 2008), yields as many descriptions as there are viewers. It is probably safe to expect that an attempt to compute *any* task-independent description of the scene with pretensions to completeness, universality, and objectivity would be misguided.

All this is probably just as well, for a reason that has to do with the failure of the other assumption behind Marr's answer to the fundamental question of vision. According to this second assumption, determining what is where and thereby constructing a representation of the world in all its 3D geometrical detail is actually going to serve some useful purpose for the seer — a notion that seems dubious, given that such a representation would need to be subjected to a detailed and far from trivial interpretation, thereby placing the seer back on square one, rather than being of any immediate use.

Textbook treatments of vision often start with a figure showing an array of numbers and an invitation to perform "object recognition" on it, the array being a numerical representation of an intensity image of some scene. But a full 3D geometrical reconstruction of a scene, listing the reflectance, position, and orientation of its every surface element, would be equally opaque with regard to whatever objects that may be present in it. At the same time, a fully annotated representation of a scene, in which all instances of all object categories from a preset list have already been flagged, would be in itself useless for finding ad hoc "objects" ("something to swat a fly with") or nameless ones ("the thing I saw on your desk the other day"), or for planning and executing any of the multitude of actions that constitute the behavioral repertoire of a human or a robot.

2.2 An alternative answer

Given that a representation that is a faithful 3D replica of the scene or a listing of all the "objects" in it, even if logically and computationally feasible, would do little to further the seer's agenda, what is it that a sense of sight would need to *make explicit* (an expression that we owe to Marr, 1982) to pay its keep? On the functional level, what an embodied, situated agent needs from its visual system is to extract from the sensorimotor data a range of affordances (Edelman and Poggio, 1989; Sloman, 1989; Poggio, 1990),[5] as defined by its circumstances, abilities, and goals. On the algorithmic level, the key computational ingredient for affordance-oriented vision is the making explicit of similarities: "To see is to form a representation of what you're looking at in terms of similarities to what you've seen on other occasions" (Edelman, 2008, p.142; Edelman and Shahbazi, 2012). On the level of brain implementation (more on which in section 5), this calls for the use of familiar mechanisms such as tuned receptive fields for similarity estimation (Edelman, 1999) and topographic maps for cue integration and sensorimotor coordination (Merker, 2007; Edelman, 2008).

One implication of this very brief (one sentence per level of explanation!) sketch of an answer to Marr's overarching question of the meaning of seeing is that all vision is purposive: what it means to see depends on what you are and what you do for a living. This realization puts an interesting perspective on the late 1980s purposive/active movement in vision (Bajcsy, 1988; Ballard et al., 1989; Aloimonos, 1990): although

[5] The concept of affordance is, of course, due to Gibson (1979), who, famously, failed to develop any appreciation of its computational meaning (Ullman, 1980).

it may seem to have petered out, it has in fact been so successful as to become the only game in town, which is why it no longer needs any special advertising. Indeed, none of the present-day computer vision methods that work (and those that do, work on the kinds of images and amounts of data that were not dreamt of in the old days) attempt anything like a reconstruction of the scene before proceeding to carry out recognition or interpretation tasks. As I shall argue next, the purposive stance has implications also for the phenomenology of vision — a cluster of questions that hitherto have largely escaped a computational treatment.

3 What does it feel like, to see, and why?

The question "What does it mean to see?" can be interpreted as calling not just for a functional (and algorithmic, and implementational) explanation of vision, but also for a phenomenological inquiry into it. Why does seeing *feel* the way it does? Why does it feel like anything at all? Is the feeling of seeing an integral part of its function? Ignoring these questions or treating them carelessly may land us in the awkward situation of having to concede that a digital camera has an inner life simply by virtue of being able to detect faces, or (much worse) that a hatful of pieces of paper, each marked with a number, is as phenomenally aware as an animal whose brain's instantaneous activity, neuron by neuron, is captured by those numbers (cf. Edelman, 2008, p.487). In this section, I outline a computational approach to the phenomenology of visual awareness that avoids such conceptual abominations.

3.1 The classical view: phenomenal vision as a representational state

Can the phenomenology of vision be explained in terms of the properties of the representation that it computes — a static snapshot of (the affordances offered by) the world? Although already in the second sentence of his book Marr insists that vision is a *process*, the rest of *Vision* makes it clear that by this he meant merely a sequence of steps that leads to a particular representation rather than something fundamentally dynamical. Had one attempted to develop a phenomenology of vision based on that account, we may speculate that it would have hinged on the nature of the final product of vision rather than on the process that leads to its completion.

The notion that phenomenal vision is static in that sense is reminiscent of a remark made by Wittgenstein (1958, II,xi), in the context of the important distinction that he drew between "seeing" or phenomenal awareness and "seeing as" or interpretation: "To interpret is to think, to do something; seeing is a state" (for a discussion, see Edelman, 2009). It is also endorsed by Smart (2004), in an encyclopedia article on the identity theory of mind: "Certainly walking in a forest, seeing the blue of the sky, the green of the trees, the red of the track, one may find it hard to believe that our qualia are merely points in a multidimensional similarity space. But perhaps that is what *it is like* (to use a phrase that can be distrusted) to be aware of a point in a multidimensional similarity space."

The appeal of the view of phenomenal awareness as a static entity — the state of a representational system — has deep roots in contemporary sensory neuroscience. Visual neuroscience, in particular, has for decades focused on studying the brain representation of stimuli presented in isolation from one another, either simply for a short time, or until the animal responded. This methodology may suffice if one assumes that the representation in question is independent of the the animal's broader behavioral context and history — an assumption that is, as I argued earlier, unwarranted on first principles, and that is also made suspect by the richness of the brain activity in the absence of what used to be called a stimulus (Vincent et al., 2007; Mason et al., 2007; Miller et al., 2009).[6]

From a computational standpoint, the static view of phenomenality is untenable for several reasons. The first of these I already hinted at: equating phenomenality with being a point in a multidimensional representation space implies that this setup is not only necessary but also sufficient for awareness. This, in turn, amounts to postulating that any system that spans a representation space, however structured and implemented, undergoes phenomenal experiences. This failure to properly narrow down the sufficient conditions for phenomenality opens the door to a particularly annoying variety of panpsychism, according to which all representational systems with the same dimensionality have the same phenomenality.

As if this predicament were not dire enough, there is also the issue of what counts as a representational "system." Are two hemispheres of a brain two systems or one? What about two brains engaged in conversation? These questions fail to fully make sense, let alone be amenable to a resolution, under the static conception of representation. By itself, a list of numbers — think of the hatful of paper notes mentioned earlier in this section, or of the "instantaneous" activities of a set of neurons — can be taken to represent anything at all, and therefore means nothing unless further processed. Static snapshots of representation spaces cannot therefore give rise to phenomenality, because phenomenal awareness cannot be a matter of outside interpretation (as in a visual representation being post-processed by another stage), but rather must be intrinsic, that is, must arise from within the system itself (Edelman and Fekete, 2012).

3.2 An alternative view: phenomenal vision as process

A proposal for a computational resolution of these issues has been recently described by Fekete and Edelman (2011). A theory of phenomenal experience based on this approach equates the phenomenality, if any, of a given system not with any of its states, but rather with a process, namely, with the evolution of the system's state over time, as directed by its dynamics. Because the minimal quintessential characteristic of

[6] Should we really be surprised that the brain is active even when not presented with a "stimulus"? The stimulus/response neuroscience is still living the Sixties' popular version of itself: the subtitle on the cover of the 1961 Science Editions paperback printing of Donald Hebb's *The Organization of Behavior* (1949) reads "Stimulus and response — and what occurs in the brain in the interval between them." I like to think that Hebb himself (as opposed to whoever was in charge of cover design for that edition) knew better than handicapping his research program in that manner.

phenomenality is discernment,[7] the system's trajectory space must be intrinsically structured so that certain classes of trajectories differ qualitatively from others — a property of the system's dynamics that can be quantified by mathematical tools such as persistent homology (Fekete, 2010).

The richness of a system's experience is thus determined jointly by the complexity of its state-space trajectories and the structure of that space, leading to a phenomenological account of vision that is inherently and inalienably dynamical (Edelman and Fekete, 2012). This account happens also to be in line with the situated, embodied, purposive take on vision advocated in the previous section. Consider, for instance, the visual experience that I have of a mug with a handle on the table in front of me. Because of the peculiarities of my embodiment (I happen to have two fully functional arms, each equipped with an end effector capable of grasping small objects), the mug affords to me at least two qualitatively distinct ways of grasping — a fact that affects my perceptual experience of it (for a couple of entry points into the vast literature documenting embodiment effects in cognition, see Anderson, 2003 and Longo, Schüür, Kammers, Tsakiris, and Haggard, 2008; embodiment with regard to consciousness is discussed by Shanahan, 2010 and reviewed by Edelman, 2011; for a range of dynamical models of phenomenal experience, see the volume edited by Edelman, Fekete, and Zach, 2012).

4 How is vision computed?

The question of the uses of vision has a direct bearing on the choice of mathematical tools required for studying visual systems — which, of course, merely reflect the mathematics of vision itself. Traditionally, the mathematics of (spatial) vision — to the extent that vision researchers considered mathematics to be important at all (Marr, 1975) — was thought to be confined to projective or affine geometry (e.g., Gibson, 1966; Hochberg, 1968). Marr's decisive argument for the overarching importance of a computational understanding of the various visual tasks precipitated a tectonic shift in this domain. Computation is now central to any theory of vision that purports to be explanatory rather than merely descriptive; and the range and the sophistication of mathematical and algorithmic tools used by vision scientists have grown considerably.

4.1 Inverse optics and regularization

If the purpose of vision is taken to be the recovery of geometry and reflectance of surfaces comprising a scene, then it is natural to treat it as inverse optics (Bertero et al., 1988). This insight reveals visual recovery to be a formally ill-posed problem, insofar as it admits multiple solutions (unless properly constrained). This, in turn, suggests that regularization can be generally used to make vision well-posed (Poggio et al., 1985).

This idea spurred many interesting developments in the research program directly inspired by Marr's

[7] Having an experiential repertoire that consists of a single quale really means being incapable of any experience.

Vision. For instance, the seemingly ad hoc assumptions required to ensure a unique solution to the correspondence problem in binocular stereopsis turned out to correspond to the regularizer term in an equation describing possible solutions. More generally, many problems in visual recovery came to be seen as instances of function approximation from examples, itself an ill-posed problem, and for the very same reason (Poggio and Girosi, 1990). This understanding, and the conceptual link that it has established between vision and the emerging science of machine learning, proved to be fruitful even as the ideas of geometrical recovery and of vision as inverse optics began losing their appeal (Edelman and Poggio, 1989).

4.2 A computational toolbox for vision

The alternative top-down — in Marr's terminology, abstract-computational — conception of vision, outlined in section 2.2, replaced the classical goal of striving for a single, general-purpose, geometrical replica of the world with a multiplicity of goals, which are dictated by the tasks at hand and which may, therefore, rely on a variety of representations and computational processes. Although its development had been motivated by other factors, this dynamic, pluralistic, top-down stance is supported also by bottom-up considerations, which belong to algorithmic and implementational levels.

Allowing bottom-up considerations to influence computational-level theorizing goes against the grain of Marr's preferred approach to integrating the levels of explanation. While noting that the levels should not be thought of as independent,[8] he held that the all-important topmost level, that of computational theory of the task, should have conceptual priority (Marr, 1982, p.337). Although this stance may be useful for understanding vision in the abstract (that is, how it *can* be done), an attempt to reverse-engineer a particular visual system (that is, how vision *is* done) is always better served by an approach that allows implementation-level findings or considerations to influence the top-level theory.

In the case of insect and vertebrate visual systems, in which the sensory front end is composed of multiple receptors (ommatidia in insects; retinal photoreceptors in vertebrates), the fundamental bottom-up constraint is that vision begins with a multidimensional set of measurements. The same goes for any computer vision that starts with a camera sensor array. In all such systems, the number of measurements produced by each snapshot of the world is typically large. In the human retina, for instance, the nominal dimensionality of the output signal that goes to the rest of the brain is about 10^6, which is the number of fibers (axons of retinal ganglion cells) in the optic nerve. Thus, mathematically speaking, vision indeed turns out to be very much about geometry — just not the projective geometry of a handful of fiducial points, but rather the geometry of manifolds embedded in multidimensional spaces (Edelman, 1999).

In itself, the claim that the subspaces of interest embedded in the measurement space are formally charac-

[8] This point has been lost on some of my fellow vision researchers. An anonymous NIH reviewer once commented on a grant proposal of mine: "...the description of the main model is full of terms such as V1, parvo, complex cell, and so on. Why should a computational model be based on so much biological detail instead of computational problems as defined by Marr?"

terized by a manifold structure is a working *assumption* (another conceptual/methodological tool borrowed from *Vision*). This assumption is, however, solidly supported — both by implementational considerations (the measurements and their subsequent transformations being carried out by smooth functions, corresponding to graded receptive fields; Edelman, 1998) and by algorithmic ones (counterfactually, were the target subspaces not low-dimensional and smooth, they could not be learned from examples; Edelman and Intrator, 2002).

The conception of vision as learning of manifolds embedded in high-dimensional spaces has been particularly successful as a basis for a theory of object and scene representation and recognition (Edelman, 1999; Edelman and Intrator, 2003; Edelman and Shahbazi, 2012), but its scope is in fact much broader. For instance, it is immediately applicable to any affordance that a visual system may be called upon to estimate, for the simple reason that such affordances can always be cast as functions from visual (and other) measurements to certain indicator or utility variables (cf. Poggio, 1990). Satisfyingly, its mathematics also connects seamlessly to the computational approach to visual phenomenology, mentioned in the previous section. Finally, it also lends itself naturally to a probabilistic treatment, thus completing a conceptual circle that connects back to Marr's earlier, tremendously insightful work in theoretical neurobiology (Marr, 1970).

4.3 The probabilistic turn and Marr's Fundamental Hypothesis

In a paper that introduced his probabilistic theory of the cerebral neocortex (a great innovation at the time), Marr included the following "Fundamental Hypothesis" (Marr, 1970, pp.150-151):

> Where instances of a particular collection of intrinsic properties (i.e., properties already diagnosed from sensory information) tend to be grouped such that if some are present, most are, then other useful properties are likely to exist which generalize over such instances. Further, properties often are grouped in this way.

Presaging a key part of Marr's later methodological stance, which stressed the reliance of vision on certain assumptions about how the world works, the Fundamental Hypothesis posits that the world is well-behaved in a probabilistic sense. Together with his recourse to Bayesian inference (normatively the right thing to do; Howson and Urbach, 1991; Edelman and Shahbazi, 2011), on which his "diagnosis theorem" is based (Marr, 1970, p.183), this insight clearly identifies Marr, well ahead of his time, as a forerunner of a conceptual revolution in theoretical neuroscience that is by now so complete that its tenets are taught to college sophomores in cognitive psychology and neurobiology.[9]

[9] An undergraduate-level textbook based on these ideas is (Edelman, 2008).

5 Where is vision in the brain?

Conceiving of vision as probabilistic and purposive may explain why so much of computer vision has now been reduced to data-driven machine learning (e.g., Malisiewicz and Efros, 2009; Choi, Lim, Torralba, and Willsky, 2010) — a development that I see as altogether positive. In comparison, the implications of this view for the understanding of the brain basis of vision are yet to be fully explored. In this section, I briefly discuss one aspect of this exploration: the shift from the corticocentric view of cognition to a broader one.

5.1 The corticocentric view

The traditional preoccupation of vision research (along with certain other areas of cognitive neuroscience) with the cerebral cortex, which has been remarked upon in the past (Parvizi, 2009), still largely persists. Although the discovery in the monkey inferotemporal cortex of neurons tuned to hands and to faces (Gross et al., 1972; Perrett et al., 1982) did surprisingly little to advance the understanding of vision (as Marr noted in *Vision*), it spurred and sustained a research program that combines an avowed interest in computational issues such as invariance with an unwillingness to consider broader computational aspects of vision, its relationships to the rest of behavior, and its roots in parts of the brain other than the cortex.

Anyone who (perhaps implicitly) holds that there is such a thing as a purely visual brain area, or that understanding the ventral and dorsal cortical processing streams in mammalian vision would settle the most important questions about it, is doing so in the face of overwhelming evidence to the contrary. Regarding the purported functional specialization of the cortex, Anderson (2010) reports that "An empirical review of 1,469 subtraction-based fMRI experiments in eleven task domains reveals that a typical cortical region is activated by tasks in nine different domains."[10] A theoretical synthesis of decades' worth of empirical work carried out by Merker (2007) shows that not just elementary visual functioning but even basic visual awareness depends more on midbrain areas, especially the superior colliculus, than on the cortex in itself. A comparative and evolutionary view offered by Balaban et al. (2010) likewise rejects corticocentrism.

5.2 A broader view

A comprehensive understanding of vision in the brain can only be achieved on the basis of the realization that vision in the wild is intimately intertwined with the rest of the brain's functions, is embodied, and is situated in the world (Anderson et al., 2012). Glimpses of such a broader understanding are already available. For instance, an integrated view of the cortical functional architecture proposes that the sensory and motor cortical domains converge to / diverge from an effective functional apex in the medial temporal lobe, consisting of the hippocampus and perirhinal and entorhinal cortices (Merker, 2004). Another kind of

[10] The domains were action execution, action inhibition, action observation, vision, audition, attention, emotion, language, mathematics, memory, and reasoning.

integration happens in the midbrain, where the layered structure of the superior colliculus brings together perception, motivation, and action (Merker, 2007). Finally, studies of action selection and executive control, working memory, and sequence processing have identified multiple parallel functional loops spanning the basal ganglia, thalamus, hippocampus, and cortex (Redgrave et al., 2011).

A computational synthesis suggests a certain division of labor among all these structures, one of the dimensions of difference being the mode of learning. The columnar cortical architecture seems to be well-suited for unsupervised learning and for the distillation of its results into a long-term memory representation of the statistics of the world, as envisioned by Marr (1970). At the same time, reinforcement learning relies on the distinct architecture of the hippocampal formation, which appears to help the rest of the brain address the structural credit assignment problem, and of the cortico-striatal loops, which are involved in temporal credit assignment and action selection (e.g., Atallah, Frank, and O'Reilly, 2004; van der Meer, Johnson, Schmitzer Torbert, and Redish, 2010). The interactions among all these cognitive processes, all of which are active in an awake brain at all times, must be understood before a complete picture of what it means to see can emerge.

6 Conclusion

Despite Marr's own fears that vision might in the end turn out to be a "bag of tricks" devoid of unifying theoretical principles (Marr, 1981), the computational study of vision, which his work was instrumental in launching, has built up an impressive record of identifying such principles and finding their signatures in the functional architecture of the brain. One meta-level principle that extends over and above these is that vision is not a general-purpose "module" whose task is to deliver a geometrically faithful replica of the world to the rest of cognition. Consequently, the study of the brain basis of vision cannot be confined to the "visual" cortex, or even to the cortical areas in general. Another one is that learning is both feasible and desirable: the world is statistically well-behaved, and so survival in it can be furthered by a combination of unsupervised extraction of visual and other sensory patterns, which correspond to objects and events, and a supervised tuning of courses of action afforded by the agent's circumstances and motives.

For a continued success of the enterprise whose goal is to make sense of how vision works, we need to preserve and promote what I believe are the most important of Marr's contributions: the license to ask big-picture questions[11] and an appreciation of the combination of mathematical skills and scholarship in psychology and neuroscience needed to frame them. If we don't try out different answers to such questions, just to see where they might lead us, the big picture of the mind will forever remain for us just a jumble of pixels.

[11] Not, however, in a grant proposal — as one of Schiller's characters remarked, "against stupidity the gods themselves contend in vain."

References

Akins, K. (1996). Of sensory systems and the ‘aboutness’ of mental states. *Journal of Philosophy XCIII*, 337–372.

Aloimonos, J. Y. (1990). Purposive and qualitative vision. In *Proc. AAAI-90 Workshop on Qualitative Vision*, San Mateo, CA, pp. 1–5. Morgan Kaufmann.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence 149*, 91–130.

Anderson, M. L. (2010). Neural re-use as a fundamental organizational principle of the brain. *Behavioral and Brain Sciences 34*, 245–266.

Anderson, M. L., M. J. Richardson, and A. Chemero (2012). Eroding the boundaries of cognition: Implications of embodiment. *Topics in Cognitive Science*. In press.

Atallah, H. E., M. J. Frank, and R. C. O’Reilly (2004). Hippocampus, cortex, and basal ganglia: Insights from computational models of complementary learning systems. *Neurobiology of Learning and Memory 82*, 253–267.

Bajcsy, R. (1988). Active perception. *Proc. IEEE 76*(8), 996–1005. Special issue on Computer Vision.

Balaban, E., S. Edelman, S. Grillner, U. Grodzinski, E. D. Jarvis, J. H. Kaas, G. Laurent, and G. Pipa (2010). Evolution of dynamic coordination. In C. von der Malsburg, W. A. Phillips, and W. Singer (Eds.), *Dynamic Coordination in the Brain: From Neurons to Mind*, Volume 5 of *Strüngmann Forum Report*, Chapter 5, pp. 59–82. Cambridge, MA: MIT Press.

Ballard, D. H., R. C. Nelson, and B. Yamauchi (1989). Animate vision. *Optic News 15*, 9–25.

Barrow, H. G. and J. M. Tenenbaum (1978). Recovering intrinsic scene characteristics from images. In A. R. Hanson and E. M. Riseman (Eds.), *Computer Vision Systems*, pp. 3–26. New York, NY: Academic Press.

Barrow, H. G. and J. M. Tenenbaum (1993). Retrospective on “Interpreting line drawings as three-dimensional surfaces”. *Artificial Intelligence 59*, 71–80.

Bertero, M., T. Poggio, and V. Torre (1988). Ill-posed problems in early vision. *Proceedings of the IEEE 76*, 869–889.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychol. Review 94*, 115–147.

Choi, M. J., J. J. Lim, A. Torralba, and A. S. Willsky (2010). Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA.

Dickinson, S., R. Bergevin, I. Biederman, J. Eklundh, R. Munck-Fairwood, A. Jain, and A. Pentland (1997). Panel report: The potential of geons for generic 3-d object recognition. *Image and Vision Computing 15*, 277–292.

Edelman, S. (1995). Vision reanimated. Unpublished manuscript; `http://kybele.psych.cornell.edu/~edelman/abstracts.html#reanimated`.

Edelman, S. (1998). Representation is representation of similarity. *Behavioral and Brain Sciences 21*, 449–498.

Edelman, S. (1999). *Representation and recognition in vision.* Cambridge, MA: MIT Press.

Edelman, S. (2002). Constraining the neural representation of the visual world. *Trends in Cognitive Sciences 6*, 125–131.

Edelman, S. (2008). *Computing the mind: how the mind really works.* New York: Oxford University Press.

Edelman, S. (2009). On what it means to see, and what we can do about it. In S. Dickinson, A. Leonardis, B. Schiele, and M. J. Tarr (Eds.), *Object Categorization: Computer and Human Vision Perspectives*, Chapter 4, pp. 69–86. Cambridge, UK: Cambridge University Press.

Edelman, S. (2011). The metaphysics of embodiment. *International Journal of Machine Consciousness 3*, 321–325. Part of a collective review of *Embodiment and the Inner Life* by M. Shanahan, Oxford University Press, 2010.

Edelman, S. and T. Fekete (2012). Being in time. In S. Edelman, T. Fekete, and N. Zach (Eds.), *Being in Time: Dynamical Models of Phenomenal Experience*, pp. 81–94. John Benjamins.

Edelman, S., T. Fekete, and N. Zach (Eds.) (2012). *Being in Time: Dynamical Models of Phenomenal Experience*. Amsterdam: John Benjamins.

Edelman, S. and N. Intrator (2002). Models of perceptual learning. In M. Fahle and T. Poggio (Eds.), *Perceptual learning*, pp. 337–353. MIT Press.

Edelman, S. and N. Intrator (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science 27*, 73–109.

Edelman, S. and T. Poggio (1989). Representations in high-level vision: reassessing the inverse optics paradigm. In *Proc. DARPA Image Understanding Workshop*, San Mateo, CA, pp. 944–949. Morgan Kaufman.

Edelman, S. and R. Shahbazi (2011). Survival in a world of probable objects. *Behavioral and Brain Sciences 34*, 197–198. A commentary on *Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition* by Jones & Love.

Edelman, S. and R. Shahbazi (2012). Renewing the respect for similarity. Submitted.

Fekete, T. (2010). Representational systems. *Minds and Machines 20*, 69–101.

Fekete, T. and S. Edelman (2011). Towards a computational theory of experience. *Consciousness and Cognition 20*, 807–827.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.

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

Gross, C. G., C. E. Rocha-Miranda, and D. B. Bender (1972). Visual properties of cells in inferotemporal cortex of the macaque. *J. Neurophysiol. 35*, 96–111.

Hebb, D. O. (1949). *The organization of behavior*. Wiley.

Hochberg, J. E. (1968). *Perception*. Englewood Cliffs, NJ: Prentice-Hall.

Howson, C. and P. Urbach (1991). Bayesian reasoning in science. *Nature 350*, 371–374.

Hubel, D. H. and T. N. Wiesel (1959). Receptive fields of single neurons in the cat's striate cortex. *J. Physiol. 148*, 574–591.

Hubel, D. H. and T. N. Wiesel (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. London 195*, 215–243.

Longo, M. R., F. Schüür, M. P. M. Kammers, M. Tsakiris, and P. Haggard (2008). What is embodiment? A psychometric approach. *Cognition 107*, 978–998.

Malisiewicz, T. and A. A. Efros (2009). Beyond categories: the visual Memex model for reasoning about object relationships. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta (Eds.), *Proc. 22nd Neural Information Processing Systems Conference (NIPS)*, pp. 1222–1230.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London B 176*, 161–234.

Marr, D. (1975). Approaches to biological information processing. *Science 190*, 875–876.

Marr, D. (1981). Artificial intelligence: a personal view. In J. Haugeland (Ed.), *Mind Design*, Chapter 4, pp. 129–142. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Marr, D. and H. K. Nishihara (1978). Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B 200*, 269–294.

Marr, D. and T. Poggio (1977). From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull. 15*, 470–488.

Mason, M. F., M. I. Norton, J. D. Van Horn, D. M. Wegner, S. T. Grafton, and C. N. Macrae (2007). Wandering minds: the default network and stimulus-independent thought. *Science 315*, 393–395.

Merker, B. (2004). Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex 40*, 559–576.

Merker, B. (2007). Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behavioral and Brain Sciences 30*, 63–81.

Miller, K. J., K. E. Weaver, and J. G. Ojemann (2009). Direct electrophysiological measurement of human default network areas. *Proceedings of the National Academy of Science 106*, 12174–12177.

Parvizi, J. (2009). Corticocentric myopia: old bias in new cognitive sciences. *Trends in Cognitive Sciences 13*, 354–359.

Perrett, D. I., E. T. Rolls, and W. Caan (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res. 47*, 329–342.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology LV*, 899–910.

Poggio, T. and F. Girosi (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science 247*, 978–982.

Poggio, T., V. Torre, and C. Koch (1985). Computational vision and regularization theory. *Nature 317*, 314–319.

Redgrave, P., N. Vautrelle, and J. N. J. Reynolds (2011). Functional properties of the basal ganglia's reentrant loop architecture: selection and reinforcement. *Neuroscience 198*, 138–151.

Reynolds, J. R. and R. Desimone (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron 24*, 19–29.

Russell, B., A. Torralba, K. Murphy, and W. T. Freeman (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision 77*, 157–173.

Shanahan, M. (2010). *Embodiment and the Inner Life*. New York, NY: Oxford University Press.

Sloman, A. (1983). Image interpretation: The way ahead? In O. J. Braddick and A. C. Sleigh (Eds.), *Physical and Biological Processing of Images*, Springer Series in Information Sciences, pp. 380–401. Berlin Heidelberg New York: Springer Verlag.

Sloman, A. (1989). On designing a visual system (towards a Gibsonian computational model of vision). *J. of Experimental and Theoretical Artificial Intelligence 1*, 289–337.

Smart, J. J. C. (2004). The identity theory of mind. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Stanford University. Available online at http://plato.stanford.edu/archives/fall2004/entries/mind-identity/.

Ullman, S. (1980). Against direct perception. *Behavioral and Brain Sciences 3*, 373–416.

Vaina, L. M. (Ed.) (1991). *From the retina to the neocortex: selected papers of David Marr*. Boston, MA: Birkhauser.

van der Meer, M. A., A. Johnson, N. C. Schmitzer Torbert, and A. D. Redish (2010). Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron 67*, 25–32.

Vincent, J. L., G. H. Patel, M. D. Fox, A. Z. Snyder, J. T. Baker, D. C. V. Essen, J. M. Zempel, L. H. Snyder, M. Corbetta, and M. E. Raichle (2007). Intrinsic functional architecture in the anaesthetized monkey brain. *Nature 447*, 83–89.

Wittgenstein, L. (1958). *Philosophical Investigations* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall. Translated by G. E. M. Anscombe.