TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 13 (2021) 309–328

ISSN: 1756-8765 online
DOI: 10.1111/tops.12527

This article is part of the topic "2018 Rumelhart Prize Issue Honoring Michael K. Tanenhaus," John C. Trueswell (Topic Editor). For a full listing of topic papers, see https://onlinelibrary.wiley.com/toc/17568765/2021/13/2.

The Hierarchical Evolution in Human Vision Modeling

Dana H. Ballard, Ruohan Zhang

Department of Computer Science, The University of Texas at Austin

Received 13 March 2019; received in revised form 22 February 2021; accepted 22 February 2021

Abstract

Computational models of primate vision took a significant advance with David Marr's tripartite separation of the vision enterprise into the problem formulation, algorithm, and neural implementation; however, many subsequent parallel developments in robotics and modeling greatly refined the algorithm descriptions into very distinct levels that complement each other. This review traces the time course of these developments and shows how the current perspective evolved to have its alternative internal hierarchical organization.

Keywords: Marr's paradigm; Active vision; Visual routines; Human gaze behaviors; Reinforcement learning; Bayes hierarchies

1. Introduction

The computational approach of the brain's realization vision dates from Marr's seminal theorizing that trudged the effort into problem specification, algorithm, and biological implementation. This organization decoupled the functional model of experimental observations from a complete account of all its neural details, which were extraordinarily complex. For the first time, computational models that addressed crucial abstract details of visual problems were formalized (Marr, 1982). However, in the 40 years since, the algorithmic approach to vision modeling has undergone elaborations into separate abstraction levels. One result is a

Correspondence should be sent to Dana H. Ballard, Department of Computer Science, The University of Texas at Austin, Austin, TX 78712. E-mail: danab@utexas.edu

Table 1
Functional levels for the study of the neuroscience of behavior

Level	Description	Timescale
Neural	Dynamics of neural circuits	20–300 ms
Embodiment	Primitive behaviors defined by fast sensorimotor co-ordination such as a fixation and pickup of an object	300–1,000 ms
Awareness	Use of simulation to modify behavioral descriptions	10 s

triage into three distinct levels, as shown in Table 1 (Ballard, 2015). A *neural level* encompasses Marr's original functional level. Its models primarily respect low-level abstractions of the visual cortical anatomy. An *embodiment level* recognizes the embodiment of vision in an active agent that utilizes visuomotor behaviors. Models at this level include image acquisition and actions via an abstract motor system. The use of embodiment here is defined by the grounding of the body as emphasized by many authors (Ballard, Hayhoe, Pook, & Rao, 1997; Clark, 2008; Noë, 2009). An *awareness level* models the behavior by an explicit conscious agent. Models at this level include instructions comprehension of audio instruction and planning (Dehaene, 2014; Graziano, 2013).

While all three levels ultimately have a neural realization, the last two are distinguished by having less dependence on neural descriptions, analogous to how silicon computers use separate computational levels. Thus, embodiment models feature the dynamics of physical systems, and awareness models feature abstract symbolical descriptions. This paper reviews these developments, situating them on the historical timeline.

2. The functional neural level: The Marr paradigm

The neural investigation of human vision saw a pivotal advance with Hubel and Wiesel's studies of edge orientation retinotopic maps in the striate cortex (Hubel & Wiesel, 1962). Retinotopic maps of other basic image features such as color, velocity, direction selectivity, and stereo were identified soon after. But how could these features be computed? The answer awaited Marr and Poggio's (1976) seminal paper to introduce the advantage of pursuing a computational account of random dot stereo to introduce a functional approach. The calculation had to be done at least in the striate cortex as it was the first map that had cells with binocular receptive fields. The elegant algorithm maximized binocular matches that had similar depths. Its success ushered in the acceptance of an abstract algorithmic approach to account for biological observations.

Besides stereo, striate cortex recordings revealed several retinotopic maps of properties such as optic flow, color, and shape from shading that were characterized as a "two and a half dimensional sketch" (Marr, 1982) and "intrinsic images" (Barrow, Tenenbaum, Hanson, & Riseman, 1978), which seemed to have common abstract functional constraints. These retinotopic two-dimensional property images turned out to have two parameters per property, but only one fundamental equation relating the image's gray-level pixels to the parameters.

For example, a retinotopic optic flow image measures the local x and y velocity vectors at each point in an image (Horn & Schunck, 1981). The basic equation was that for every point in an image $I(x, y, t)$, the total variation had to be zero, that is, $\frac{d(I(u,v,t,))}{dt} = 0$, which could be expanded to

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y^v} = \frac{\partial I}{\partial t}.$$

Thus, at each point in the image, there was the above equation with two unknowns, $u(x, y, t)$, $v(x, y, t)$. To make the system of equations well posed, typically spatial smoothness, in which the changes in u and v would vary smoothly across the image, was invoked to provide the needed additional constraints.

The impact of these results cannot be overestimated. Observable properties of cells in early cortex maps were understandable in terms of local constraints that potentially could be compatible with local networks.

It was time to introduce the image's observer. What constraints would drag one's attention to one item in an image over another? The proposed answer was hugely influenced by the work of Triesman (Treisman & Gelade, 1980), whose work suggested that a combination of local image features might be an attentional stimulus. Objects defined with simple features such as color "popped out." Thus, a green object was quickly located in a cloud of red objects, whereas combinations of features seemed to require that objects be inspected individually, taking much more time. Thus, the concept of image saliency was born with the seminal paper by Koch and Ullman (1987). While the hindsight of much later research shows that most situations require knowing the observer's goals to account for looking patterns (Tatler, Hayhoe, Land, & Ballard, 2011; Triesch, Ballard, Hayhoe, & Sullivan, 2003), in many cases saliency provides an informative choice of useful image features (Bruce & Tsotsos, 2006; Itti & Koch, 2001; Koch & Ullman, 1987).

Like saliency, ever exacting primal sketch models made creative assumptions to achieve their results, but it burned out that at the theorizing Marr level, their assumptions have had to be revised. Most of all, the smoothness constraint was challenging for natural images. Such images were rife with discontinuities that could defeat that objective function guiding the recovery of the properties from Marr's two and a half degree sketch (two spatial coordinates and a physical property).

Another issue was the models' locality. The decision was made to focus on the computations that could be modeled locally within the striate cortex retinotopic map. While it was known from the anatomy that there were connections from more abstract cortical maps, Van Essen et al.'s definitive hierarchical anatomy maps were not published until 1992 (Van Essen, Anderson, & Felleman, 1992). Thus, a research strategy was to constrain the models to have locally striate cortex connections.

Finally, a general assumption of the time was to emphasize two-dimensional image inputs. This stance was widely held by the general vision research community at the time. The thinking was that two-dimensional images were simpler than three-dimensional images and thus should be a priority. After all, the animal data could be obtained from anesthetized preparations, and the additional complication of an awake behaving animal posed new technical

difficulties. The research impetus was to get the issues with two-dimensional images solved first and then move on to the more complicated time-varying three dimensions of the natural world.

3. The embodiment level: Active vision

Bajcsy (1988) was the first to recognize that seeing was an active process and that the computer vision enterprise had to move on:

> Most past and present work in machine perception has involved extensive static analysis of passively sampled data.
>
> However, it should be axiomatic that perception is not passive but active. Perceptual activity is exploratory, probing, searching; percepts do not simply fall onto sensors as rain falls onto ground.

This quote has at least two new ideas. One is that perceptions are not passive but are ordered up by an active process of the perceiver. Another is that rather than a complication, active perception provides additional useful constraints for the perceiver. The ability to move the head provides a straightforward relationship between depth and optic flow. The depth cues from binocular vergence movements program the eye's horizontal rotational vestibulo-ocular reflex appropriately.

The Bajcsy laboratory's binocular computer-controlled moving camera system shown in Fig. 1 was an instant success when it was unveiled at the IEEE conference in Traverse City, Michigan, in 1985. Thus, in one instant, the computational vision enterprise became embedded into the world of the animate perceiver where the vision properties of the cortex had to deal with a world regularly in motion. Besides an algorithmic level confined to the two-dimensional venue of abstract neural circuits, a new dimension was added, engendered by including the perceiving agent's sensorimotor apparatus.

The new level required new hardware. To access this world required moving cameras and sufficient computational cycles to keep up with the image video inputs, and quickly other designs emerged at Rochester (Brown, 1988) and KTH in Sweden (Pahlavan & Eklundh, 1992) had real-time processing. Rochester's binocular cameras were mounted on a platform that gave each camera independent yaw control, and had independent pitch control. The system had the capability of making saccadic fixations at a considerable fraction of 700° per second—the speed of human saccades, and could make vergence and pursuit movements. The real-time processing system, shown in Fig. 2, used *DataCube* video processing circuit boards. The boards contained a series of circuits, each containing programmable functionalities. A given control strategy typically used several such boards, which could be interconnected. For example, to compute kinetic depth, separate boards would compute vertical and horizontal image gradients. Knowing the fixation point allows inferring the sign of each point's depth, as points behind the fixation point move in a direction opposite to the camera motion and vice

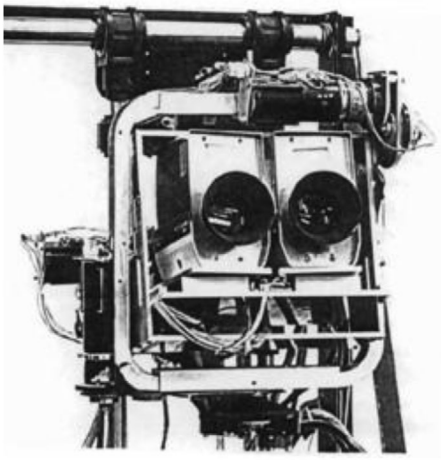

Fig. 1. Bajcsy's camera system. A joint project between the University of Pennsylvania's computer science and mechanical engineering departments. It was the first to combine pitch and yaw degrees of freedom in a binocular system. The known movements provided egocentric parametric information that simplified many visual computations.

versa for points in front of the fixation point. Thus, a large subset of image processing could take advantage of the new constraint of self-motion using simpler and more robust algorithms to recover physical properties from image pixels.

The active perception movement led to the construction of several laboratories' robot heads and the incorporation of the human repertoire of specific movements of saccades, vergence, and pursuit for refined scientific models as well as improved robot vision systems. However, the robot heads had a much more significant impact than real-time dynamic image processing. Early vision theorizing had focused on a "bottom-up" perspective, wherein processing started at the retina and proceeded through brain regions. In contrast, the robots introduced "top-down" models with a focus moved to the agenda of the seeing agent. Nonetheless, from the perspective of a complete account of human visual capabilities, something is still missing.

The use of vision to solve such problems had been started earlier. This component was a focus for Just and Carpenter (1976) and Kosslyn (1980), but it was up to Ullman to make the explicit connection that this focus had been missing from the Marr paradigm (Richards & Ullman, 1987; Ullman, 1987). Two of his examples appear in Fig. 3. Both of these represent a problem that cannot be solved without calculation.

While the problem-solving approach to vision was essential and timely, the advent of robot heads added a crucial component of platforms that could test constructive models. Subsequently, the visual routines model of human visual computation was taken up in both the algorithms and implementation levels by several laboratories (Aloimonos, 1990; Findlay, Findlay, & Gilchrist, 2003; Pahlavan, Uhlin, & Eklundh, 1993). In one paradigmatic example, Swain and Ballard (1991) showed that explicitly making a cortical parietal-temporal pathway

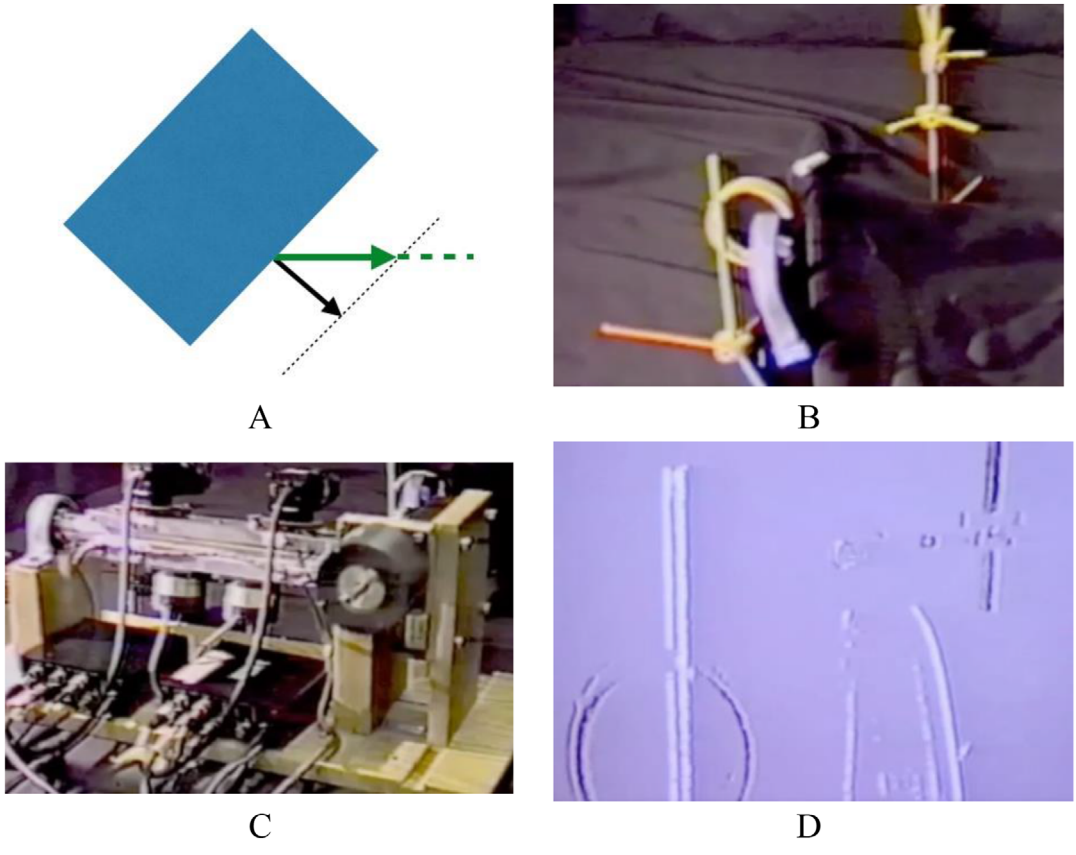

Fig. 2. Kinetic depth demonstration. (a) Basic constraints. A basic horizontal motion of the head is shown by a green arrow. The fundamental image constraint: Motion perpendicular to the edge can be measured. Putting this projection together with a motor efference copy signal allows the true motion to be recovered. (b) Laboratory table scene with multicolored objects and a suspended piece of white chalk. (c) Robot head. (d) Kinetic signal recovered. During translated motion, the camera tracks the chalk with the result that points in front of the fixated chalk, which can be colored white, and the point behind can be colored black.

distinction (Mishkin, Ungerleider, & Macko, 1983) led to significantly simpler algorithms for both object localization and identification.

3.1. Back to level I: Cortical hierarchies; the Bayes and neural routines

Meanwhile, the discovery of regularities in the connection between cortical maps led to a complex of hierarchies (Van Essen et al., 1992) that revealed that adjacent maps had point-to-point connections. This revolutionary discovery was soon followed by mathematical formalisms to explain striate cortex receptive fields in terms of coding economies (Olshausen & Field, 1997), followed by an interpretation of the connections between these maps wherein cells' receptive field codes can be interpreted in terms of their ability to predict the cell

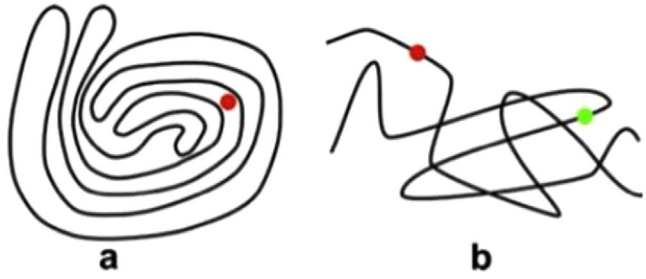

Fig. 3. Ullman's visual routines examples. (a) The question is to determine whether or not the red dot is inside or outside of the curve. This problem can be solved by counting the intersections on a line from the dot to the image edge. (b) The question is whether or not the green and red dots are on the same curve. This problem can be solved by tracing the green dot along its curve and seeing if it meets the red dot. The essential is that these problems are outside of the realm of computing primal sketch properties.

responses of maps below in the hierarchy (Rao & Ballard, 1999). The error in this prediction has another critical use as a learning signal. Such a signal can be used to adjust receptive field synapses. Another critical feature of this learning signal is that it can also be applied throughout the cortical hierarchy.

Given that these hierarchies have characteristic connectivity patterns, it opens up the possibility of another circuit interpretation, which is that the cortex is a vast compendium of behavioral experience. Thus, we can invoke Bayes theorem to express a likelihood as

$$P(W|S) = P(S|W)P(W)$$

with the following identifications:

1. W: the state of the world
2. S: the sensory data.

On the right-hand side, the first term is the likelihood, which records the probability of seeing the data given the state of the world, and the second term records the probability of a world state. Typically, there is some constraint that guides the choice of probabilities for a world state that allows the likelihoods to be computed. A paradigmatic example comes from a motion illusion by Weiss, Simoncelli, and Adelson (2002) that has a ready Bayesian explanation as shown in Fig. 4.

A horizontally moving rhombus is rendered at different contrasts. In the space of optic flow values, the constraints from the likelihoods of the linear edges appear as lines, but the low contrast case results in blurring. As a consequence, when combined with a prior for flows that prefers slow motions, the result of the Bayes prediction is for downward motion in the blurred case, congruent with the human perception.

The influence of the Bayes interpretations cannot be overestimated. In the first place, the perceptions do not have to be solely linked to sensory data but can be economically and flexibly expressed as the product of the likelihood and prior (Doya, 2007). Second, Bayesian coding can be factored into all the levels in the cortical hierarchy. Without this perspective, the

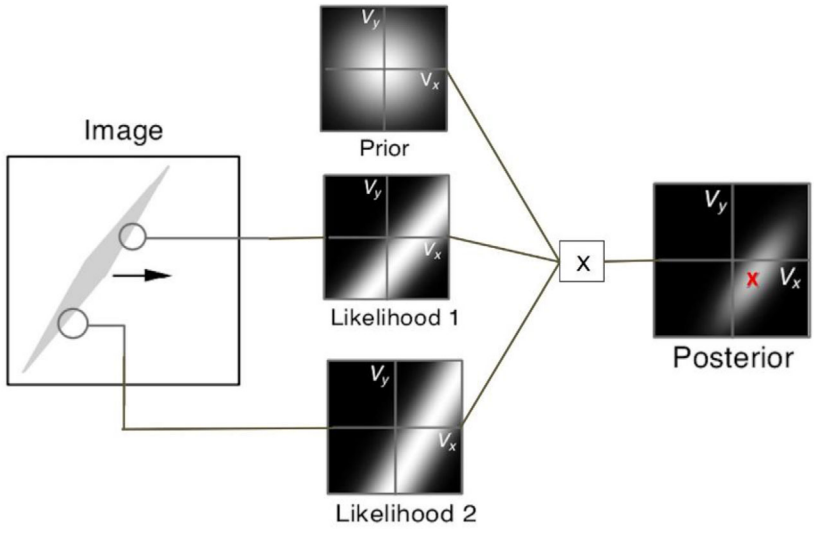

Fig. 4. Bayesian priors affect the perception of optic flow (*Vx, Vy*). A horizontally moving rhombus with (*V*0, 0) is perceived differently depending on its contrast. For strong contrast, the motion would be seen as veridical, but for weak contrast shown, the perceived motion has a downward component indicated by the red X. This difference can be accounted for if the perceiving system has a preference for slow motion, indicated in the figure by the distribution of optic flow vectors centered about zero motion. If local estimates of flow are sharp, they can overcome the prior bias, but if low contrast estimates are noisy, the maximum product of priors and likelihoods has a downward shift indicated by the red X in the posterior. (Redrawn from Weiss et al. [2002].)

burden of the incoming image from the retina is to have an exacting code commensurate with the phenomenon of seeing; with it, the code just has to be sufficient to activate an appropriate network of priors.

3.2. Visual routines at the neural level

Another development in the spirit was the demonstration of attentional effects at the neural level (Ito & Gilbert, 1999; Moran & Desimone, 1985; Tsotsos, 2011). These works had been the demonstration of more refined characterizations of visual routines, both in animals and humans. In a tracing task (shown in Fig. 5), Roelfsema showed evidence that a monkey solved a connectedness problem by mentally tracing a curve (Roelfsema, Khayat, & Spekreijse, 2003). By arranging to record from a single-oriented cell on the path of a traced curve, he showed that the time course of the cell's elevated firing was consistent with the tracing's process.

4. The awareness level: Agenda-based use of gaze behaviors

In introducing this level, we need to disambiguate it from level II, which can be similar in several ways. In the latter, the focus is on elucidating how visuomotor coordination is realized

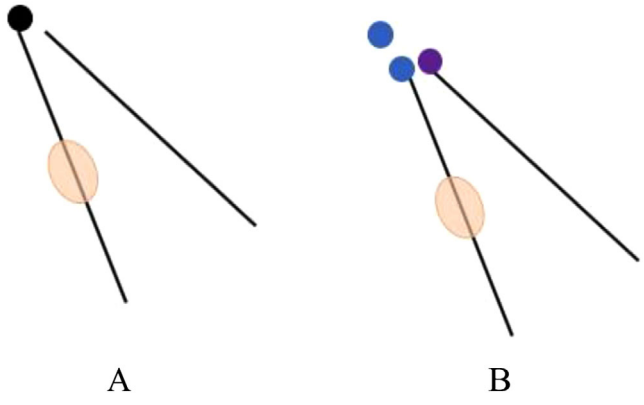

Fig. 5. Rolfsema's visual routines demonstration. (a) A monkey is trained to hold fixation on a single black dot. Next, two lines appear. The monkey must make a saccade to the end of the line that contacts the fixation dots. In one of many trials, the experimenters are recording from a cell that has an appropriately oriented receptive field on the target line. The result is that spike activation is delayed for an interval consistent with mental tracing being used to determining the saccadic target. (b) In a more difficult task, the monkey must pick the line that has the same color stub as the color of the fixation point. The hypothesis was that the time to resolve the correct target should take longer as two problems—color matching and tracing—have to be solved. The result is a cell's activation is delayed appropriately.

with no emphasis on cognition. Level II is exemplified by subconscious behaviors like one can tune out while driving a car on a familiar route and have no memory of the interval.

In contrast, level III uses behaviors that could become subconscious, but a prime focus is on a subject's ability to plan, incorporate verbal instructions, and respond to on-off cues.

Head-free video-based eye tracking: In primate vision, a central cue is the use of gaze fixation. Gaze control is impressively sophisticated, with six separate systems working in concert. All of these systems subserve the goal of stabilizing the gaze on the moving head, but of these, the use of saccades to fixate stationary targets and pursuit to track moving targets have a particular association with the subject's aims.

Early eye trackers were not head free and cumbersome to use. The dual Purkinje trackers required the subject to use a lab-mounted bite bar to stabilize the head during measurements (Van Rensbergen & De Troy, 1993). Land's free head video system was much better but required the gaze to be recovered frame by frame using image post-processing, reminiscent of Yarbus's much earlier cornea-mounted mirror system (Yarbus, 1967). Meanwhile, yet another technical advance made its appearance in the form of light-weight head-mounted eye tracking. The importance of individual fixations had been long appreciated since Yarbus, but early head-restrained trackers were significantly limiting experimental questions. Modern systems track the pupil via video to get a reliable eye-in-head signal to obtain an accuracy of 1–3 visual degrees.

Yarbus's experiments showed that the gaze was intimately involved in the scene for the viewer's cognitive goals in scanning a scene and question answering. This new capability

allowed this capability to deconstruct the progress of real-time visual analyses into their component tasks in hand-eye coordination tasks (Findlay et al., 2003; Gray & Boehm-Davis, 2000; Kowler, Anderson, Dosher, & Blaser, 1995). One of the first of these showed the coordination of a block copying task. Subjects shown a target pattern of a handful of colored blocks on a screen had to make a copy by selecting blocks from a reservoir with a cursor, drag them to a construction area and arrange them in an appropriate copy of the target. Fixations were used to direct the copying stages (Ballard, Hayhoe, Li, & Whitehead, 1992). Similar levels of coordination were observed in the sequential tapping of arrangements of dots (Epelboim et al., 1995).

Land showed predictable visual coordination, hand-eye coordination in driving (Land & Lee, 1994), and intercepting cricket pitches (Land & McLeod, 2000) along with the more involved tasks in tea making, followed by Hayhoe's analysis of sandwich making (Land & Hayhoe, 2001). Copying a pattern of blocks provided a moment-by-moment appreciation of the memory management of the process (Ballard, Hayhoe, & Pelz, 1995).

Language and vision showed to synchronize: Outside of the efforts in visual-motor coordination, the study of language had been limited to single word presentations and reading. Tanenhaus was quick to recognize the new capability's possibilities for situated language comprehension in the real world. In displays containing multiple objects, the actions asked of subjects could be studied in the context of the meaning of the objects and their phonemic content. His paper with Spivey linking fixations with question answering was ground-breaking (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In responding to directions, fixations were in lock-step with an utterance on a millisecond scale, a completely new finding. Situated behavior had been shown to have a comprehensive visual-audio embedding, but the tight temporal synchronization pointed to an underlying synchronization of neural computations across modalities.

This new paradigm had a huge impact and spawned many related research questions. In infant language learning, Chen and Smith showed that progress was intimately connected to cues from the caregiver given in real-world settings (Yu, Ballard, & Aslin, 2005; Yu & Smith, 2012).

As for human studies during the early 2000s, the arrival of head-mounted displays incorporating eye trackers was game-changing. Developed under a National Center for Research Resources (NCRR) facility at the University of Rochester, the innovation allowed human studies of gaze choices in virtual environments. Experimenters could study the information extracted by a fixation by manipulating the image at video frame rates under computer control. Studies like that of Droll and Hayhoe (2007) showed that fixations could result in particular results. In a virtual reality block sorting task when the color of a held block was changed just before a sorting decision, subjects used the remembered color acquired during an earlier pickup rather than the changed color. This kind of result is very much in resonance with the Bayesian principles discussed earlier. Since the color had been inventoried at the beginning of the task using a visual routine, there was not a reason to check again as the prior probability that objects typically do not change their color. Whether or not the image information is recorded is modulated by the observer's goals.

Table 2
Reinforcement learning learns a policy that specifies how to choose an action from any given state. Policies may be probabilistic. The main feature of the computation is that it computes actions based on expected discounted rewards

State	A state might be a phase in the preparation in a cooking recipe
Action	An action takes the chef from one phase to another
Transition function	A transition function describes the probability of transiting from a phase to another given a choice of action.

4.1. Mainstream reinforcement learning

Given that agenda-driven experience is ubiquitous, the next logical question is how such routines get programmed in the first place. The complete answer to this question is still open, but the general abstract answer is that routines are learned by reward (Kaelbling, Littman, & Moore, 1996). The neurobiological evidence for this in primates stems from Schultz (1998), which has been formally associated with reward by Schultz, Dayan, and Montague (1997).

Reinforcement learning works with states, actions, transition functions, and rewards, as shown in a cursory example in Table 2. The modern formalism of reinforcement algorithms was pioneered by Barto and Sutton, summarized in a classic text that has been recently updated (Sutton & Barto, 2018).

Thus, progress in reinforcement learning has brought us to the point that dovetails with intellectual progress in cognitive science. The original questions of Marr in the 1970s were confined to the early cortical areas, and neural models paved the way for Bajczy's active vision that incorporated the extra degrees of freedom of the circumscribed vision system. The insights that the addition brought allowed a progression to even more formulations of embodied cognition that integrated gaze and motor actions into cognitive models. At this point, this wave of progress has brought us to the point that we can start to revisit and update Newell's program of a "unified theory of cognition," and start to think what that model would look like, given all the legacy insights learned from vision studies.

Formal reinforcement learning seems like the outline of an answer to how the brain programs itself, but many details are unresolved. The biggest problem is that realistic reinforcement learning models scale exponentially with the size of the state space of the system. Handling large state space reinforcement models has been ameliorated with advances in deep learning (LeCun, Bengio, & Hinton, 2015), and reinforcement learning has been sped up with improvements such as episodes (Botvinick et al., 2019), but given exponential scaling problems remain.

Without a way forward, there seems no way of adapting reinforcement as a model in cognition. Thus, research is examining ways to decompose a more extensive problem into a set of modules specialized in solving more tractable smaller problems.

One view is that the next frontier needs to address how many streams of thought can be active at one time. This issue surfaced 20 years ago in neuroscience as the "binding problem," denoting the difficulty of parsing the interconnected neural substrate into usefully

distinct components, but has lain fallow since Von der Malsburg (1999). However, more recently, given a more cognitive context, interest in multiplexing has resurfaced (Panzeri, Brunel, Logothetis, & Kayser, 2010). Alternate motor plans are simultaneously active before choosing between them (Cisek, 2012). This result resonates with another from the Tanenhaus paradigm, where it was shown that alternate object choices with names that had overlapping phonemes interfered on the millisecond level (Salverda et al., 2007). One interpretation of these speedy responses is that the names were co-active at the neural level as observed by Cisek. Even more remarkable, analyses of cortical recordings suggest that in a population of spikes, only 14% are used to represent significant task features, suggesting that the remaining 86% are being used for other processes (Kobak et al., 2016). Also, there is a growing general interest in neuroscience-based multiplexing proposals (Akam & Kullmann, 2014).

Given the experimental evidence for multiple simultaneous reinforcing "threads," researchers have focused on ways to compress the exponential space needed to account for behavior. One long-standing approach has been that of Sutton's *Options* (Precup, Sutton, & Singh, 1998; Stolle & Precup, 2002; Sutton, Precup, & Singh, 1999), which associates the result of a program with its initial states of a lengthy coding of the internal state descriptions of how to bring about the result. Another is that of constraining the state space into parsimonious descriptions of small *modules*. This tack defines modules a priori (Sprague, Ballard, & Robinson, 2007), as depicted abstractly in Fig. 6.

It turns out that modules defined this way have several important properties that speak directly to cognitive concerns so that they are touched upon here.

Multiprocessing has a long-standing tradition of study in psychology in the form of dual tasks, but the influence of module reinforcement learning formalisms in multitasking is significant. The modules approach can serve as the underpinning for a "cognitive operating system" that can manage ongoing behaviors. The central assumption is that at any instant, the behavioral demands can be factored into a small set of active subsets. An example could be cooking the components of a meal. The different components on a stove require temporal attention as to their progress.

While a small number of modules can be simultaneously active, the composition of the modules can be changed quickly, allowing for a rich tapestry of different modules to be active over time. The modal fixation interval of 200–300 ms provides a logical time constant for module switching. The temporal agility of the operating system context allows complex behaviors to be addressed. In the example of driving, the car radio can be turned off in a situation of congested traffic.

A particular specialization for modules occurs when the action taken is shared by these modules. A typical case is accounting for the allocation of gaze for sets of modules that have spatially separated targets of interest. In this setting, multiple modules are co-active (Sprague et al., 2007). Their state is drifting in time with an attendant increase in loss of reward as the policies cannot be as good in the uncertain state estimates. Given that two modules' states cannot be simultaneously updated, which one should get the gaze? A driving study (Johnson, Sullivan, Hayhoe, & Ballard, 2014) showed that for each module, the expected loss in reward for not looking could be computed, and the module with the highest expected loss should

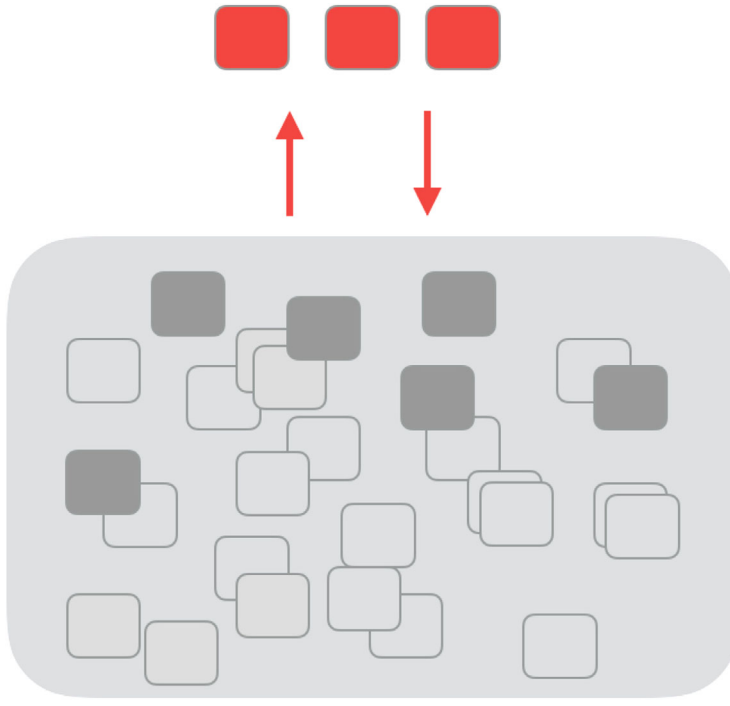

Fig. 6. Modules hypothesis. A general abstract module of behavior consists of enormous numbers of self-contained sensorimotor behaviors that are used in small numbers of compatible subsets to avoid neural cross-talk. Most often, these behaviors would be automatic, but if they have special activation constraints, they would require working memory. Thus, working memory would be analogous to the computer science concept of "threads," the state needed to keep track of an ongoing computational process. Modules can be activated and deactivated on a 300-ms timescale to track environmental reward opportunities. Shown in dark gray are modules that would be appropriate candidates for the current context.

have its state updated with a fixation. The model accounted for the distribution of human gaze intervals among multiple simultaneous behaviors, as shown in Fig. 7.

The dynamics of behavior management are the most significant advantage of the modules operating system approach, but here are two of several others.

One is that there is a ready solution for inverse reinforcement learning (Lopes, Melo, & Montesano, 2009; Neu & Szepesva´ri, 2007; Ng & Russell, 2000). A significant component of learned experience comes from watching others. Given sequences of state-action pairs, the task of inverse reinforcement learning is to recover the reward distributions that were used to generate them. For many situations, one can assume that the watcher has similar basic modules as a demonstrator but lacks the knowledge of how they are used in combination. The straightforward module formulation of the problem assumes the demonstrator is using a weighted combination of modules, and the problem for the learner is that of determining these weights from samples of the demonstrator's behavior. This approach, together with a Bayesian formulation, results in exposing the weights in a way that they are easily recovered.

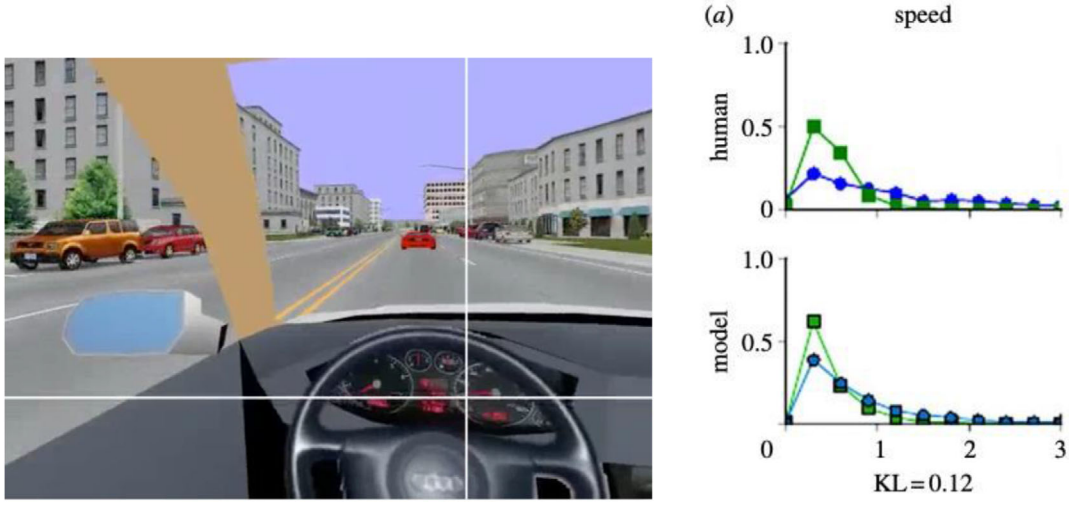

Fig. 7. A driving setting in virtual reality allows the study of multiplexing. A subject tasked with following a red car and keeping the speed at 30 mph shares a gaze between the speedometer and the followed car. At the instant shown gaze, indicated by the cross-hairs, is reading the location of the dial indicator for speed. Successfully executing both tasks can be achieved if separate modules for them are active and continually use gaze to update their state estimates of speed and car location.

In a sidewalk navigation setting, Rothkopf and Ballard (2013) and Zhang et al. (2018b) were able to show that this works well for a navigation task replete with obstacles, but the method would generalize to any modularizable setting.

For the other example of the power of the modular setting, consider the issue the brain has in calculating the reward values of all its active sets of modules. Given that behaviors are selected on the basis of reward, how can the appropriate reward values be calculated? The modules formalization has an answer to this question if one assumes that active modules have access to the total reward earned at every instant but are not given explicit information as to how to divide it up. In this setting at each instant, a module r_i has two estimates of its reward. One is its running estimate, and another is the immediate reward handed out minus the estimates of the other modules in the currently active module cabal. Thus, the estimate can be updated as the weighted sum of these two using the weight parameter β:

$$\hat{r}_i := (1-\beta)\,\hat{r}_i + \beta \left(\sum_j r_j - \sum_{j!=i} \hat{r}_j \right)$$

This formula is very mathematically well behaved and takes particular advantage of the expectation that over time modules will be active in different combinations (Rothkopf & Ballard, 2010). The methodology is elegant and awaits substantive verification, but the importance of being able to bootstrap credit assignment online should not be underestimated.

In summary, the module formalism is one approach to define cognitive control at the abstract level. It is not unique and has its modern-day competitors such as Sutton et al. (1999)

and its temporally distant relative such as Newell and Allen (1994) and Anderson (1996) that are more programming language based. Nonetheless, these approaches share the goal of defining the execution of cognition at an abstract level.

5. The future

This review has described a progression toward understanding human vision. At the start, research chose to be limited to two-dimensional image models, but over the intervening years progressed to a three-dimensional behavioral setting where models now can have an interface with the motor system and the agent's goals. Research has continued to evolve a modeling perspective of hierarchies of algorithms to deal with brain issues at its different levels in Newell's style (Ballard, 2015; Newell & Allen, 1994).

For addressing the hardest open questions, Marr's abstract logical methodologies are still active and significant. However, the biggest revision to the Marr paradigm is the floodgate of techniques to characterize the function of neurons and their networks. This progress is two-way: The abstract models can be described at a lower level, and the low-level data can inform the abstract structures. Thus, the focus of the discussion will be the issues situated in a living brain and visual system.

One trend will continue to be incomplete systems. While the beginning of visual models focused on a compartmental approach in characterizing early vision computations, in current research, active vision is commonplace, with complete complements of multiple gaze modes integrated into the hand, head, and body movement systems that produce robust solutions to complete sensorimotor behaviors. These systems are becoming large scale (Jain, Xiong, & Grauman, 2017), but the focus of this embodiment is automatic behaviors. Such automatic behaviors are valuable for their owners, owing to their economic overheads, and they are also implicitly the focus for deep learning methods.

Another trend will be toward models of the idiosyncratic flexibility in human learning. Human brains arguably use mostly autonomous behaviors but also can exhibit flexible behavior to handle unique situations. The overall system that handles these cases have been usefully denoted as *awareness* by Graziano (2013). One crucially important question for the brain is defining the interface between automatic and modifiable behaviors. Graziano's distinction can be interpreted hierarchically; awareness denotes a topmost level description, whereas attention is reserved for the embodied program that carries this description out, and vision, being our most complex sense, has ready parsing into each of these two levels.

A proper discussion of awareness is beyond this review, but a bare outline will serve to describe the issue. Novel features of behavior are extracted by the brain's hippocampus and amygdala systems and somehow incorporated into the brain's cortical memory during sleep. Although we understand the subsystems at this level and more, the implementation of the models that can exhibit this functionality is just beginning, but the visual system would be expected to play a pivotal role (Muller, Chavane, Reynolds, & Sejnowski, 2018). Awareness is also starting to get a discussion on a Marr-like level. Its importance has been given increased focus by prominent theorists, such as Dehaene (2014) and Graziano (2013).

Another easy prediction is the continued importance of Tanenhaus has on innovative demonstrations that language's rapid decision-making and production moves in a lock-step coupling with the perceptual-motor apparatus. This research program has made an enormous impact that is far from over, but this new venue brings up the relevance of modules (Ballard, Kit, Rothkopf, & Sullivan, 2013; Sprague et al., 2007; Van Seijen et al., 2017; Zhang et al., 2018b). Their promise can be illustrated with the example of *multiplexing* that brain computation modelers are starting to tackle. Silicon computers routinely take advantage of their high processing speeds to produce real-time multiprocessing by sharing small processing instants between multiple programs. It seems implausible that the brain would not have a solution as well, even if its space-time form were very different.

A crucial open multiplexing issue is how to choose good behaviors from a plethora of alternatives. That humans are addressing this problem can be appreciated from the study of "change blindness," the phenomenon that humans can be oblivious to significant changes to their environment (Simons & Levin, 1997). Nevertheless, incompletely opposite to its moniker, change blindness is a hugely important filter that ignores vast numbers of unimportant or irrelevant distractors. This everyday human venue wherein a relatively small number of reward possibilities are chosen from the vast number of alternatives stands in contrast to some situations wherein *all* exigencies have to be handled.

So given this massive number of choices, how do we determine what events can interrupt our cognitive agenda? This question is unanswered, but an ingenious suggestion comes from a study by Roth et al. (2016). A rat running on a treadmill is subjected to two different sets of visual stimuli. In one set, the stimuli are appropriate for the rat's motion. In another set, the visual motion is disjunct. It turns out the different cases are routed to the cortex using separate thalamic neural "buses." What the rat can predict is one bus, and what that the rat cannot is another. The fact that this distinction is implemented into the rat's anatomy is the ultimate indication of its central importance. Thus, a hugely valuable insight is that things that can interrupt cognition may be those associated with being unpredictable and are handled by a particular anatomical pathway that registers things that cannot be predicted *with respect to the current cognitive agenda*. The study by Triesch et al. (2003) also gives a huge insight: The task relevance of a changed sensory feature influences if it is noticed.

The most straightforward prediction for vision is the continuing importance of the deep learning revolution (LeCun et al., 2015; Schmidhuber, 2015; Sejnowski, 2018), with its evolving focus on human capabilities (Merel, Botvinick, & Wayne, 2019; Mnih et al., 2015). Deep learning allows much more exacting real-time cognitive tests to be explored with brain-like architectures. These efforts are usually computer-based, exploiting the million to billion-time advantages of silicon architectures. However, a trend is to cross this divide and use human eye movement data and action choices to inform deep learning (Zhang et al., 2018a).

Although deep learning is proven to be extraordinarily powerful, many open questions are specifying an interface with the human neural system. The effort to a formative interface has been already started (DiCarlo, Zoccolan, & Rust, 2012; Mhaskar & Poggio, 2016) but still need to tackle the details of neural computation at millisecond levels (Fries, 2015; Maris, Fries, & van Ede, 2016; Panzeri et al., 2010; Vinck &Bosman, 2016) and rationalize the various forebrain oscillations. It has now been established that the forebrain uses meshed

oscillations at least four frequencies: theta, alpha, beta, and gamma in the course of its computation (Landau, Schreyer, Pelt, & Fries, 2015; Vinck, Lima, Womelsdorf, Oostenveld, & Singer, 2010; Zheng & Colgin, 2015). A coherent account of their interactions will be necessary for understanding at the neural level as it is an essential substrate for understanding human cognition.

Acknowledgments

The paper benefited greatly from comments from Jochen Triesch and Wayne Gray. Its preparation was supported by NIH grant R01 RR09283.

References

Akam, T., & Kullmann, D. M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience*, *15*(2), 111.

Aloimonos, J. (1990). Purposive and qualitative active vision. *10th International Conference on Pattern Recognition* (vol. 1, pp. 346–360). Los Alamitos, CA: IEEE Computer Society Press.

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*(8), 966–1005.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80.

Ballard, D., Kit, D., Rothkopf, C. A., & Sullivan, B. (2013). A hierarchical modular architecture for embodied cognition. *Multisensory Research*, *26*, 177–204.

Ballard, D. H. (2015). *Brain computation as hierarchical abstraction*. Cambridge, MA: MIT Press.

Ballard, D. H., Hayhoe, M. M., Li, F., & Whitehead, S. D. (1992). Hand-eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *337*(1281), 331–339.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–742.

Barrow, H., Tenenbaum, J., Hanson, A., & Riseman, E. (1978). Recovering intrinsic scene characteristics. In A. Hanson & E. Riseman (Eds.). *Computer Vision Systems* (Vol. 2, pp. 3–26). New York: Academic Press.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408–422.

Brown, C. M. (1988). The rochester robot. Technical report, Rochester University, NY, Department of Computer Science.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), *Advances in neural information processing systems* (pp. 155–162). Cambridge, MA: MIT Press.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, *22*(6), 927–936.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York, Penguin.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Doya, K. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1352.

Epelboim, J., Steinman, R. M., Kowler, E., Edwards, M., Pizlo, Z., Erkelens, C. J., & Collewijn, H. (1995). The function of visual search and memory in sequential looking tasks. *Vision Research*, *35*(23-24), 3401–3422.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.

Fries, P. (2015). Rhythms for cognition: Communication through coherence. *Neuron*, *88*(1), 220–235.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrate- gies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied*, *6*(4), 322

Graziano, M. S. A. (2013). *Consciousness and the social brain*. New York: Oxford University Press.

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1-3), 185–203.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional archi- tecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Ito, M., & Gilbert, C. D. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron*, *22*(3), 593–604.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194.

Jain, S. D., Xiong, Bo, & Grauman, K. (2017). Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2117–2126). Los Alamitos, CA: IEEE Computer Society.

Johnson, L., Sullivan, B., Hayhoe, M., & Ballard, D. (2014). Predicting human visuomotor be- haviour in a driving task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1636), 20130044.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., & Machens, C. K. (2016). Demixed principal component analysis of neural population data. *Elife*, *5*, e10989.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of intelligence* (pp. 115–141). Dordrecht: Springer.

Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, *35*(13), 1897–1916.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25-26), 3559–3565.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, *369*(6483), 742.

Land, M. F., & McLeod, P. (2000). From eye movements to actions: How batsmen hit the ball. *Nature Neuroscience*, *3*(12), 1340.

Landau, A. N., Schreyer, H. M., Pelt, S. V., & Fries, P. (2015). Distributed attention is implemented through theta-rhythmic gamma modulation. *Current Biology*, *25*(17), 2332–2337.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lopes, M., Melo, F., & Montesano, L. (2009). Active learning for reward estimation in inverse reinforcement learning. In W. Buntine, M. Grobelnik, D. Mladenić & J. Shawe-Taylor (Eds.), *Machine Learning and Knowledge Discovery in Databases*, ECML PKDD 2009. Lecture Notes in Computer Science (vol. 5782, pp. 31–46). Berlin, Heidelberg: Springer.

Maris, E., Fries, P., & van Ede, F. (2016). Diverse phase relations among neuronal rhythms and their potential function. *Trends in Neurosciences*, *39*(2), 86–99.

Marr, D. (1982). *Vision: A computational approach*. Cambridge, MA: MIT Press.

Marr, D., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, *194*(4262), 283–287.

Merel, J., Botvinick, M., & Wayne, G. (2019). Hierarchical motor control in mammals and machines. *Nature Communications*, *10*(1), 1–12.

Mhaskar, H. N., & Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, *14*(06), 829–848.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Belle- Mare, M. G., & Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Frontiers in Cognitive Neuroscience*, *229*, 342–345.

Muller, L., Chavane, F., Reynolds, J., & Sejnowski, T. J. (2018). Cortical travelling waves: Mechanisms and computational principles. *Nature Reviews Neuroscience*, *19*(5), 255.

Neu, G., & Szepesvári, C. (2007). Apprenticeship learning using inverse reinforcement learn ing and gradient methods. *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, (pp. 295–302). Corvallis, OR: AUAI Press.

Newell, & Allen (1994). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. *17th International Conference on Machine Learning* (pp. 663–670). San Francisco, CA: Morgan Kaufmann.

Noë, A. (2009). *Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness*. New York: Hill and Wang (Macmillan).

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, *37*(23), 3311–3325.

Pahlavan, K., & Eklundh, J. -O. (1992). A head-eye systemanalysis and design. *CVGIP: Image Understanding*, *56*(1), 41–56.

Pahlavan, K., Uhlin, T., & Eklundh, J. -O. (1993). Active vision as a methodology. In Y. Aloimonos (Ed.) *Active perception* (pp. 19–46). Hillsdale, NJ: Lawrence Erlbaum Associates.

Panzeri, S., Brunel, N., Logothetis, N. K., & Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. *Trends in Neurosciences*, *33*(3), 111–120.

Precup, D., Sutton, R. S., & Singh, S. (1998). Theoretical results on reinforcement learning with temporally abstract options. *10th European conference on machine learning* (pp. 382–393). New York: Springer.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79.

Richards, W., & Ullman, S. (1987). *Image Understanding 1985–1986*. Norwood, NJ: Ablex.

Roelfsema, P. R., Khayat, P. S., & Spekreijse, H. (2003). Subtask sequencing in the primary visual cortex. *Proceedings of the National Academy of Sciences*, *100*(9), 5467–5472.

Roth, M. M., Dahmen, J. C., Muir, D. R., Imhof, F., Martini, F. J., & Hofer, S. B. (2016). Thalamic nuclei convey diverse contextual information to layer 1 of visual cortex. *Nature Neuroscience*, *19*(2), 299.

Rothkopf, C. A., & Ballard, D. (2010). Credit assignment in multiple goal embodied visuomotor behavior. *Frontiers in Psychology*, *1*, 173.

Rothkopf, C. A., & Ballard, D. H. (2013). Modular inverse reinforcement learning for visuomotor behavior. *Biological Cybernetics*, *107*(4), 477–490.

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition*, *105*(2), 466–476.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Sejnowski, T. J. (2018). *The deep learning revolution*. Cambridge, MA: MIT Press.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267.

Sprague, N., Ballard, D., & Robinson, Al (2007). Modeling embodied visual behaviors. *ACM Transactions on Applied Perception (TAP)*, *4*(2), 11.

Stolle, M., & Precup, D. (2002). Learning options in reinforcement learning. *International Symposium on abstraction, reformulation, and approximation* (pp. 212–223). Berlin: Springer.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1-2), 181–211.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 5.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, *3*(1), 9.

Tsotsos, J. K. (2011). *A computational perspective on visual attention*. Cambridge, MA: MIT Press.

Ullman, S. (1987). Visual routines. In M. A. Fischler & O. Firschein (Eds.) *Readings in computer vision* (pp. 298–328). Los Altos, CA: Morgan Kaufmann.

Van Essen, D. C., Anderson, C. H., & Felleman, D. J. (1992). Information processing in the primate visual system: An integrated systems perspective. *Science*, *255*(5043), 419–423.

Van Rensbergen, J., & De Troy, A. (1993). A reference guide for the Leuven dual-PC controlled Purkinje eye-tracking system (Psychological Report No. 145). Leuven, Belgium: Laboratory of Experimental Psychology, University of Leuven.

Van Seijen, H., Fatemi, M., Romoff, J., Laroche, R., Barnes, T., & Tsang, J. (2017). Hybrid reward architecture for reinforcement learning. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach & R. Fergus (Eds.) *Advances in Neural Information Processing Systems* (pp. 5392–5402). Red Hook, NY: Curran Associates, Inc.

Vinck, M., & Bosman, C. A. (2016). More gamma more predictions: Gamma-synchronization as a key mechanism for efficient integration of classical receptive field inputs with surround predictions. *Frontiers in Systems Neuroscience*, *10*, 35.

Vinck, M., Lima, B., Womelsdorf, T., Oostenveld, R., & Singer, W. (2010). Sergio Neuen- schwander, and Pascal Fries. Gamma-phase shifting in awake monkey visual cortex. *Journal of Neuroscience*, *30*(4), 1250–1257.

Von der Malsburg, C. (1999). The what and why of binding: The modelers perspective. *Neuron*, *24*(1), 95–104.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598.

Yarbus, I. (1967). *Eye movements and vision*. New York: Plenum Press.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29*(6), 961–1005.

Zhang, R., Liu, Z., Zhang, L., Whritner, J. A., Muller, K. S., Hayhoe, M. M., & Ballard, D. H. (2018a). Agil: Learning attention from human for visuomotor tasks. In V. Ferrari, M. Hebert, C. Sminchisescu & Y. Weiss (Eds.) *Computer Vision–ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*. (pp. 663–679). Cham: Springer.

Zhang, R., Zhang, S., Tong, M. H., Cui, Y., Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2018b). Modeling sensory-motor decisions in natural behavior. *PLoS Computational Biology*, *14*(10), e1006518.

Zheng, C., & Colgin, L. L. (2015). Beta and gamma rhythms go with the flow. *Neuron*, *85*(2), 236–237.