YALE JOURNAL OF BIOLOGY AND MEDICINE 92 (2019), pp.127-137.

PERSPECTIVES

YALE JOURNAL OF BIOLOGY AND MEDICINE
YJBM
CELEBRATING 90 YEARS 1928–2018

Attention: The Messy Reality

John K. Tsotsos*

Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada

The human capability to attend has been both considered as easy and as impossible to understand by philosophers and scientists through the centuries. Much has been written by brain, cognitive, and philosophical scientists trying to explain attention as it applies to sensory and reasoning processes, let alone consciousness. It has been only in the last few decades that computational scientists have entered the picture adding a new language with which to express attentional behavior and function. This new perspective has produced some progress to the centuries-old goal, but there is still far to go. Although a central belief in many scientific disciplines has been to seek a unifying explanatory principle for natural observations, it may be that we need to put this aside as it applies to attention and accept the fact that attention is really an integrated set of mechanisms, too messy to cleanly and parsimoniously express with a single principle. These mechanisms are claimed to be critical to enable functional generalization of brain processes and thus an integrative perspective is important. Here we present first steps towards a theoretical and algorithmic view on how the many different attentional mechanisms may be deployed, coordinated, synchronized, and effectively utilized. A hierarchy of dynamically defined closed-loop control processes is proposed, each with its own optimization objective, which is extensible to multiple layers. Although mostly speculative, simulation and experimental work support important components.

INTRODUCTION

It is almost universal to regard attention as the facility that permits an agent, human or machine, to give its resources priority for processing relevant stimuli while ignoring the irrelevant. The reality of how this might manifest itself throughout all the forms of perceptual and cognitive processes possessed by humans, however, is not as clear. Here we examine this reality in order to highlight the manner by which attentional processes may be controlled. Before proceeding, it is useful to be clear about how attention is defined here. Attention has the goal of matching current task to processing resources by manipulating the brain's processing machinery to preferentially process expected stimuli while ignoring the irrelevant, yet monitoring all stimuli in order to deal with the unexpected. Specifically for vision, attentional mechanisms provide for a dynamic tuning of visual processing in reaction to the task, goals, and input of the moment [1,2]. Although the hope is to provide an attentional control framework that is general, most of this presentation will focus on vision.

There is a large literature on attentional control; here, only some highlights will be mentioned as they relate to our main point. Kahneman was very concerned with the optimality of attentional allocation [3]. He advocated that continuous, coherent, serially organized behavior is an important determinant of effective attentional con-

*To whom all correspondence should be addressed: John K. Tsotsos, Department of Electrical Engineering and Computer Science, York University, 4700 Keele St., Toronto, ON Canada M3P 1J3; Tel: 416-736-2100; x70135, Email: tsotsos@cse.yorku.ca.

†Abbreviations: ST, Selective Tuning; CP, Cognitive Programs; STAR, Selective Tuning Attentive Reference model; STAR-AX, STAR's Attention Executive; STAR-FC, Star's Fixation Control.

Keywords: vision, attention, executive control, Selective Tuning

trol. Allport [4] is much closer to our view, saying it is pointless to focus on the brain's locus of selection and on which processes do and do not require attention. He claims that attentional functions are of very many different kinds, serving a great range of computational functions. Egeth and Yantis [5] suggest that the key issues are: control of attention by top-down (goal-directed) and bottom-up (stimulus-driven) processes; representational basis for visual selection, including how much attention can be said to be location or object based; time course of attention as it is directed to one stimulus after another. Corbetta and Shuman [6] consider the brain regions involved and conclude that partially segregated networks carry out different attentional functions: the intraparietal and superior frontal cortex prepare and apply goal-directed selection for stimuli and responses; and, the temporoparietal cortex and inferior frontal cortex specialized for detection of behaviorally relevant stimuli. Rossi and colleagues [7] claim that frontal and parietal cortices are involved in generating top-down control signals for attentive switching, which may then be fed back to visual processing areas. The prefrontal cortex in particular plays a critical role in the ability to switch attentional control on the basis of changing task demands. For Miller and Buschman [8], visual attention may be focused via a frontoparietal network acting on the visual cortex. These network interactions may be regulated via rhythmic oscillations. The brain may operate discretely with pulses of activity routing packets of information.

None of these works and most others do not, however, provide a mechanistic view of attention, in the sense of Brown [9]: "most ask what brain regions are active during attentive processes or what networks are active instead of what mechanisms are necessary to reproduce the essential functions and activity patterns in an attentive system." On the other hand, cognitive architectures, reviewed in Kotseruba & Tsotsos [10] do propose specific mechanisms. However, as that review demonstrates most take a very narrow view of attention and there is little agreement on how attentional control is accomplished. What follows, therefore, is a proposal of some steps towards how control may be mechanistically achieved. To begin, we need to be explicit about the attentional characteristics under consideration. The literature describes many phenomena connected to attention directly or indirectly. The following list gives the basic attentional phenomena (adapted from [1] where pointers to groundbreaking papers for each can be found):

Alerting: The ability to process, identify, and move attention to priority signals.

Binding: The process that correctly combines visual features to provide a unified representation of an object.

Covert Attention: Attention to a stimulus in the visual field without eye movements.

Disengage Attention: The generation of signals that release attention from one focus and prepare for a shift.

Endogenous Influences: Endogenous influence is an internally generated signal used for directing attention, including domain knowledge or task instructions.

Engage Attention: The actions needed to fixate a stimulus whether covertly or overtly.

Executive Control: The system that coordinates the elements into a coherent unit that responds correctly to task and environmental demands including selecting, ordering, initiating, monitoring, and terminating functions.

Exogenous Influences: Exogenous influence is due to an external stimulus and contributes to control of gaze direction in a reflexive manner.

Inhibition of Return: A bias against returning attention to previously attended location or object.

Neural Modulation: Attention changes baseline firing rates and firing patterns of neurons for attended stimuli.

Overt Attention: Also known as **Orienting**, the action of orienting the body, head, and eyes to foveate a stimulus in the 3D world.

Preattentive Features: The extraction of visual features from stimulus patterns perhaps biased by task demands.

Priming: Priming is the general process by which task instructions or world knowledge prepares the visual system for input. **Cueing** is an instance of priming; perception is speeded with a correct cue, whether by location, feature, or complete stimulus.

Recognition: The process of interpreting an attended stimulus, facilitated by attention.

Salience/Conspicuity: The overall contrast of the stimulus at a particular location with respect to its surround.

Search: The process that scans the candidate stimuli for detection or other tasks among the many possible locations and features in cluttered scenes.

Selection: The process of choosing one element of the stimulus over the remainder. Selection can be over locations, over features, for objects, over time, and for behavioral responses, or even combinations of these.

Shift Attention: The actions involved in moving an attentional fixation from its current to its new point of fixation.

Update Fixation History: The process by which the system keeps track of what has been seen and processed which are used in decisions of what to fixate and when.

Visual Working Memory: Attention seems necessary to select the stimuli and level of their interpretation stored in visual working memory. Working memory may impact subsequent perceptual actions.

This list is likely incomplete and as likely, many readers will disagree with one or more of its elements.

Where there should be more agreement is that there are many different manifestations of visual attentive behavior that have been reported (see [11-13]). Regardless of specific differences regarding this list, the reader should immediately be concerned about how it can be that these are connected and coordinated so that they lead to observed human attentive behavior.

ATTENTIONAL MECHANISMS AND THEIR COORDINATION

There are a few basic realities that play a major role towards the key argument of this paper. Any brain or behavioral process takes time. Each computation takes time, each transfer of information from one representation to the next takes time because information must travel over some neural distance, each motor action takes time, and so on. Each behavior has a specific time course and different visual sub-tasks, each require different actions, different timings of actions, different durations, etc. It feels immediate to us when we look at a photo of a single face and detect that it is our mother, yet it will necessarily take much longer to find that same face in a crowd. The former task may not require eye movements yet the latter may need many, each requiring about 250ms to set up and execute. The number of different kinds of visual behaviors humans can execute seems unbounded and the variability of both time and success of execution equally variable. It could be that exactly the same machinery is used in exactly the same manner for all behaviors, and that for difficult cases, that machinery simply takes longer to "settle" on an answer, perhaps using some kind of slow information accumulation. This possibility can be rejected because, if true, we would not observe the changes seen in neural and brain imaging experiments (which motivated the perspectives in [3-10]). It is rather clear that there is a sequence of actions orchestrated to achieve a particular behavior as all the authors cited earlier agree.

Consider the simple act of detecting a face in an image. Suppose you are a subject in a perceptual experiment. You are told that you will be shown a series of pictures and in advance, will be asked a question about the picture to which you should answer as quickly as possible. The picture will be shown as long as you have not replied, but once you reply, it will be removed. What is the sequence of actions within your visual system that are in play for this task? First of all, the initial instructions you receive set you up to expect a picture and to receive a question. The question sets up your visual system to expect something; say it's your mother's face. Whatever might be irrelevant to this task might be suppressed or ignored by your visual system. From an efficiency point of view, the irrelevant should be ignored to conserve resources. Then the picture appears and you detect its location and move your gaze to fixate the picture, perhaps in the center as a default. The resulting image is processed by your expectant visual system. It is a picture of a single face on a background depicting a dining room. Presumably, some decision process would realize that the neurons that represent your mother's face are strongly firing and would then signal you to respond positively to the question posed. The picture is taken away and your gaze returns to perhaps something else in anticipation of the next picture. Another question is asked in advance of the next image; say it's to detect your father's face, and you re-tune your visual system to prepare for this. This re-tuning might instruct the system to no longer expect your mother's face, either in terms of its size, features or location and impose the characteristics of your father's face as the relevant ones, suppressing others. The next picture is shown and again you orient your gaze to it. This time, it is not a single face on a known background but a group of people in an outdoor scene where one of them might be your father. The same process, which has now been primed as the one to use for detection and decision-making is no longer the right one. The face is now much smaller, and is present in a conflicting background because there are many faces and they all share some similarities simply by being faces. You decide to begin a search but this requires you first to re-tune your visual system to process the smaller faces. You locate a face and then move your eyes to it and scrutinize the small features to check if that face could be your father's. You move from one to the next until you look at them all. If you cannot find your father, it is likely you might go back and check some of the faces just in case you made a mistake, perhaps the lighting or viewpoint led to mis-leading cues. Eventually you will decide whether or not you detect your father's face. This is a much longer process and one that requires eye movements and all the processing they entail. Although this description is an abstract one, and certainly cannot be said to include all details, it suffices to argue for why some kind of controlling process is needed. The important questions are how are all these actions and decisions taken at the right time, in the right order, and monitored for their correctness?

It would not be unreasonable to suppose that you know how to search a picture for a target because over time you have learned the algorithm. Such an algorithm for a complex visual behavior has been termed a Cognitive Program [14] or Visual Routine [15]. It might be that you have learned thousands of such cognitive programs (CP†) and you have them stored in memory, quickly deploying the right one at the right time, dynamically parameterized for the task. These CPs would provide an encoding of algorithms for visual behaviors, including attentional actions. Many elements from the list of attentional mechanisms are easily evident in these examples:

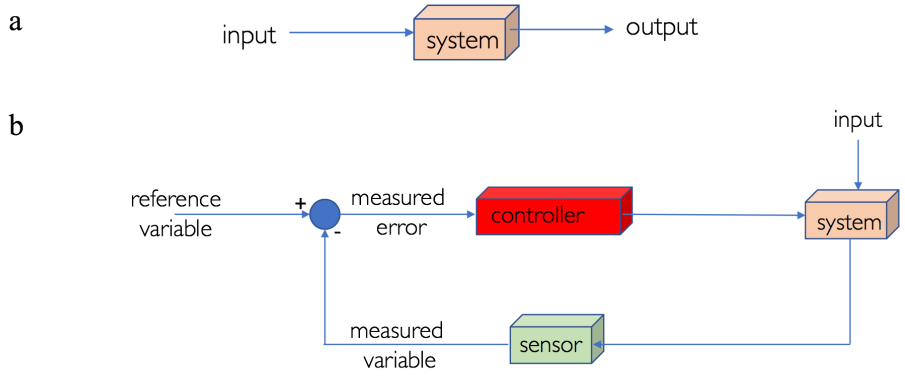

Figure 1. Classic control models. **a**) Open-loop control includes no method for monitoring or altering the system's output. **b**) The standard closed-loop control model where the controller provides signals to the system that will bring it closer to a reference state. The system's state is sensed continually and compared to the reference in order to determine these adjustments.

covert attention, disengage attention, engage attention, inhibition of return, neural modulation, overt attention, priming, recognition, search, selection, shift attention, and visual working memory. Generally, the timings of these actions need synchronization and coordination. The CP algorithm for a particular visual behavior represents an ordered and parameterized sequence of actions comprised of these attentive elements, in addition to other computations.

STAR-AX

STAR is the cognitive architecture [14] based on the Selective Tuning (ST) model of visual attention [1] and AX is its Attentional Executive sub-component. The key element of STAR-AX is that it impacts a visual processing architecture that permits control: the vast majority of proposals and systems for visual processing are fixed, rigid and unchanging. They have little or no dynamic character. Certainly in the era of deep learning as the dominant methodology for vision systems, the previous assertions may seem odd. However, it is important to note that the dynamic character required is of a particular type. It needs to enable significant changes of the visual processing hierarchy in a moment to moment basis that reflect the nature of the task being performed and the visual context in which it is being performed. It does not refer to learning the basic system nor to learning over a period of time by adjusting that basic system as more data is acquired. It refers to dynamic tuning for the current task. STAR is designed to execute the behavior algorithms represented by CPs in such a dynamic manner.

ST incorporates a layered architecture for visual processing where the computations within each layer can be dynamically affected or attentively modulated by an executive. Examples of such modulations include priming for a preferred task stimulus in some spatial location as well as in feature dimension [16], top-down suppression of interfering contextual stimuli within a receptive field [17], suppression of irrelevant or already processed inputs [18], etc. (see also [1]). This is in stark contrast to vision proposals old (*e.g.*, [19]) and new (*e.g.*, [20,21]) that do not permit any attentive control. They typically provide no functionality even for eye movements.

Within ST, let us consider what sort of control STAR-AX might embody. Classical control methods can be considered either as open-loop or closed-loop. In open loop control, the control action from the controller is independent of the process output whereas in closed loop control, the control action from the controller is dependent on feedback from the process in the form of the value of some process variable (think of the difference between heating a house without or with a thermostat). See Figure 1.

Feedback control seeks to maintain a prescribed relationship of one system variable to another by comparing functions of these variables and using the difference as a means of control. Open-loop control is what most current proposals for visual processing embody; but as the brief overview of attentional executive control above shows, this is not likely what is present in the brain. It remains then to determine what kind of process variable (or variables) our visual systems might be using and how control may be performed.

One thing that seems clear is that there might be more than one process variable. If one considers the

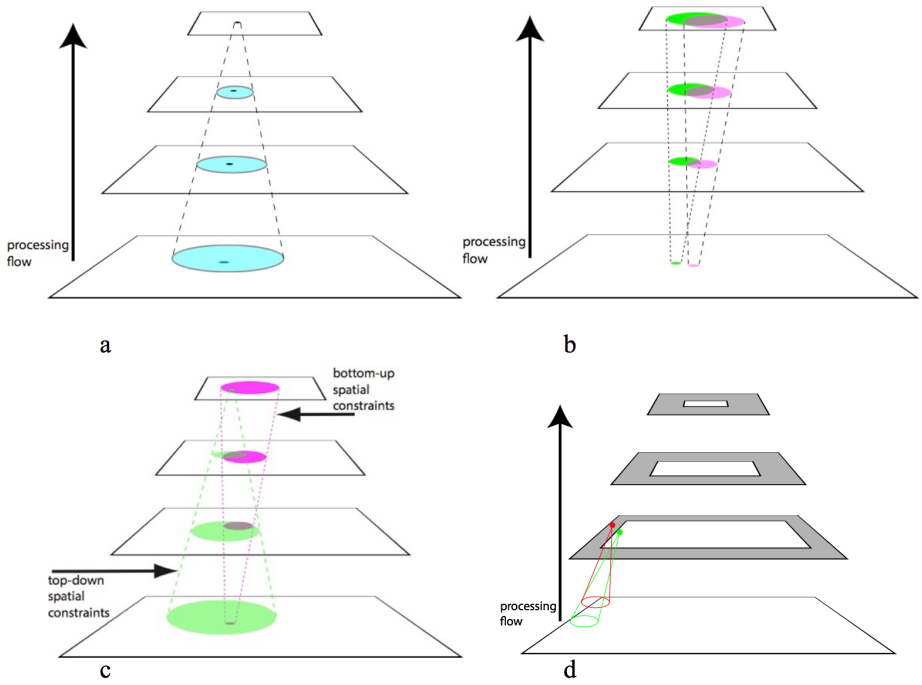

Figure 2. The breadth of problems inherent in pyramid representations. **a**) The Context Problem. A stimulus (black dot) within the receptive field of a top layer neuron, showing its spatial context defined by that receptive field. **b**). The Cross-Talk Problem. Two input stimuli activate feed-forward projections that overlap, showing the regions of overlap containing neurons that are affected by both. Those might exhibit unexpected responses with respect to their tuning profiles. **c**) The Routing Problem. Interacting top-down and bottom-up spatial search constraints are shown with the areas of overlap representing the viable search regions for best neural pathway. **d**) The Boundary Problem. The two units depicted in the second layer from the bottom illustrate how the extent of the magenta unit's receptive field is entirely within the input layer while only half of the receptive field of the green unit is within the input layer. The bottom layer represents the input and higher layers the subsequent process representation. The boundary problem forces more and more of the periphery to not have a veridical representation in higher layers of the pyramid. (Adapted from [1]).

many different flows of information and control in the brain (*e.g.*, [6,8] or proposals in [10]) it seems unlikely that a single dimension of optimization would suffice for such complex networks. Consistent with our goal of a mechanistic approach for control, we begin with Marr's computational level of analysis [19]: consider first what the nature of the problem might be and what might the nature of a solution look like.

We begin with a computational foundation for visual attention. It has long been believed, and formally proved, that the computational difficulty of vision indicates processing power much larger than what the brain provides [1,2,22,23]. One way by which the brain makes the computational problem easier is to represent visual information and its processing in a pyramid representation – a layered hierarchy of neuron arrays [22,24]. Each neuron has a receptive field centered at one location of the image and each is tuned to be selective to some feature or feature combination. The pyramid has an input layer and then each layer of processing subsequent to the input is performed by an increasingly lower spatial resolution layer. All neurons have a limited region in an image from which they receive input (feedforward converging connections in a many-to-one fashion) and a similarly sized region to which they provide input to the next processing layer (diverging feedforward connections in a one-to-many fashion). Each similarly receives feedback and lateral connections. The reason such an architecture makes the task easier is that it reduces the size of representations that need to be processed, but in doing so, sacrifices, or trades off, spatial resolution. Pyramid feedforward structure is now almost universally part of modern machine vision methods. As useful as this architecture might be, it also presents a number of issues most of which have not been widely studied (but see [25]). Some of these pyramid problems are depicted in Figure 2 (see also [1,2]).

There are three major problems that will be discussed here. The first, signal interference, can be seen in Figure 1a and b, in the former case due to local context impacting every neuron's response while in the latter case

due to intersection of the feedforward diverging connectivity patterns inherent in the architecture. The result is that although the visual hierarchy can see (encode) everything, it might not always be able to distinguish one thing from another because of the interference (see effective demonstrations in [26]). This is a direct consequence of the pyramid architecture trading off spatial resolution for a reduction in representational size. This kind of interference is common in other domains and a well-known solution is available, namely, adaptive beamforming. An adaptive beamformer is a system that performs signal processing by dynamically manipulating the combination of signals (how they interfere with each other) so that the signal strength to/from a chosen direction is enhanced while those to/from other directions are degraded (this is commonly used in cellular communication). Adaptive beamforming seeks to maximize the signal-to-interference-plus-noise ratio *S/(I+N)* [27]. In order to use this method – dynamic control of constructive and destructive interference – a localized representation is required. A distributed representation (such as one where all concepts are represented using a code only recognizable by considering a population of responses) could not be so controlled. Generally, *S*, *I,* and *N* are not known in advance for our domain. *S* is dependent on input and goals and can only be hypothesized or predicted (*e.g.*, expect to see your mother's face). *I* is dependent on context but defined by network connectivity (*e.g.*, context is whatever is in an image that is not the target of expectation, the room background in our example). *N* is dependent on context but defined by neural processes and their inherent stochasticity. Noise cannot be controlled but *I* can be controlled because the way information flows through the processing network can be manipulated. Thus, the strategy that will permit the system to maximize *S/(I+N)* is one where predictions are made as to what the signal of interest might be (is a particular patch of an image that of my mother's face?) and actions are taken to manipulate *I* so as to achieve the largest increase in the strength of the network responses corresponding to the prediction *S*. These actions include: priming the network for *S* by suppressing portions of the network that are irrelevant (when looking for my mother's face, the network units sensitive to image features related to office buildings are likely irrelevant, among others) or imposing a surround suppression within the network for candidate neurons in order to reduce any context effects. These attentional actions are basic elements of the CP for this task.

A different sort of closed-loop control process is required to solve the Routing Problem (Figure 2c). The problem arises when trying to decide which neurons and which pathways represent the best interpretation of the stimulus. The figure illustrates the extent of spatial constraints that may be used by a decision process in its search through the pyramid for the best interpretation. If the search is bottom-up — from stimulus to highest-layer neuron — then the search is constrained to the feed-forward cone outlined by the magenta lines and ovals. If the decisions are based on locally maximal neural responses, then there is nothing to prevent a bottom-up search missing the globally maximum response at the top layer (the neuron in green). It is clear that to be successful, the correct path must always go through the overlap regions shown in dark ovals. But nothing guarantees that the local maximum must lie within those overlap regions. If the search is top-down — from the globally maximum responding neuron to the stimulus — the search is constrained by the dashed lines and green regions. Only top-down search is guaranteed to correctly connect the best responding neuron at the top with its stimulus (theoretically proved in [1]). Even if the search direction is correct, there is still the possibility that its starting point, the neuron chosen at the top to which a decision process ascribes representative of the prediction *S,* might be incorrect. If correct, then as the search proceeds downwards, suppressing the irrelevant neurons, the response of that top neuron will monotonically increase, motivated by the observations of [28]. Thus, this determines the control process variable (successfully tested in [17,29]), whose measurements during the search process would satisfy the following. Let the response at time t of the neuron that represents S in network layer λ be given by $\rho_\lambda(S,t)$. The search proceeds from the top layer $\lambda = L$ to the earliest layer $\lambda = 1$. Then, if the search is proceeding successfully, for each network layer, $\rho_\lambda(S, t+1) \geq \rho_\lambda(S, t)$. This applies for S as a whole or any of its parts which would naturally be selected on a downwards network traversal. Comparing responses between layers is not appropriate since receptive field sizes differ and this would be orthogonal to the control goal.

Figure 2d shows the boundary problem, another of the issues with pyramid representations. Basically, the nature of spatially-limited convolution processing (and spatially-limited neural receptive fields) leads to a half-kernel width of undefined responses around the boundary of the visual field [1]. This effect is additive in a feedforward layered hierarchy and thus can represent a substantial portion of peripheral vision in higher layers. In current machine vision methods, the boundary problem is finessed by image padding. However, the brain does not have this luxury – there is no image padding. In fact, a more useful solution is used, namely eye movements [30,31]. This means that there must be some process to determine when the lack of veridical boundary representation is problematic and to define a new gaze fixation point that would remedy it. Moreover, there must be some process, ever vigilant to the full visual field, to determine when events within that non-veridical boundary must be attended (*e.g.*, the flashing red light in your peripheral vi-

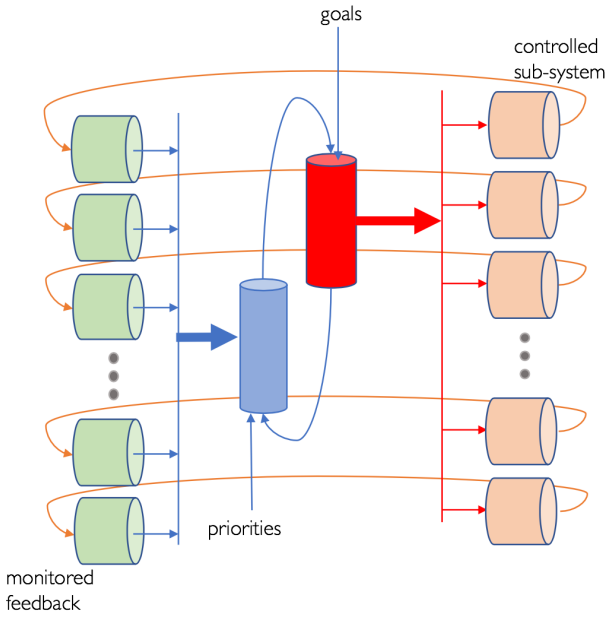

Figure 3. The closed-loop control of STAR-AX. The intent is the same as the standard closed-loop model except now there are several closed-loop controllers (horizontal loops) all tied to a higher level controller (vertical loop); it is hierarchical (with perhaps yet a higher level controller at the task level driven by a CP [14]). The color coding of the elements has similar meaning to that of Figure 1b. The reference can be dynamically set depending on task goals shown as an input to the central control loop (corresponding to the reference variable of Figure 1b, but now being dynamically set). The "orange" elements are the processes that are controlled. The "green" elements denote the measured or observed values for each of the control variables. The "blue" element corresponds to the computation of the sum of deviations from the goals; the system priorities are an input to this. The "red" element determines the values process control $\gamma(i, t)$ based on system goals.

sion). A model with effective simulation performance has been previously presented within ST [18,30,31].

The point of an eye movement is to bring the region of highest acuity to the current location of interest in a scene. S is the hypothesized image element of interest. Since the fovea, and specifically its foveola, has the highest density of cones, it should always be placed at the location of maximum interest, or if there is no particular location of interest, then at the centroid of S as a default. Let the gaze position be (x, y) in retinal coordinates and the centroid of S be (x_S, y_S). The closer the point of gaze – or the center of the foveola – is to the object centroid, the more retinal cones will fall within the object, so a controller will seek the following objective,

$$\min_{(x,y)} \left\| (x, y) - (x_S, y_S) \right\|$$

at any time t. In previous work, we presented a control loop for gaze fixation prediction whose performance is within measurement error of human sequence prediction [18,31]. A key computation used in determining the prediction is visual salience, but of a very specific kind and not simply local image conspicuity. Our model, termed STAR-FC (STAR's fixation controller) computes salience separately based on features peripherally and on objects centrally and modulates these depending on fixation history, task, and context to create a priority map upon which fixation decisions are made. The controlling loop is to move fixation through the ordered set of priority map maxima, *i.e.*, the sequence of (x_S, y_S) of interest, which reflects this control objective.

By now it might be apparent that there is an element in this description that is not found in standard control theory methodologies. Classic methods do not include the ability to change the process variable dynamically, but this is exactly what is needed here. The variable S that represents the hypothesized item of interest changes dynamically as our tasks change. Moreover, because S is a hypothesis, it might be found incorrect and replaced with a different hypothesis. As is seen above, there are potentially many process variables, one for each of the separate attention mechanisms. As a result, STAR-AX proposes a re-formulation of classic closed-loop control by including a dynamic setting of reference points. The new control loop is depicted in Figure 3.

How is overall control achieved? The relevant control literature is that which pertains to multi-objective optimal control (*e.g.*, [32]). First, optimization frameworks require some method of measuring the deviation between

observed and reference (or desired) variables. Several such variables have been presented but there certainly are more. We can thus specify this deviation as,

$$\Delta_i(K_r^i(t) - K_\gamma^i(t))$$

where Δ is an appropriate norm over the *i*-th control variable. $K_r^i(t)$ is the reference value of the *i*-th control variable at time *t*. $K_\gamma^i(t)$ is the observed value of the *i*-th control variable under the control signal *γ(i, t)*. The set of control signals is given by Γ={*γ(1, t), γ(2, t)..... γ(N, t)*}, for N control processes at time *t.* Second, we note that each control variable will not necessarily have the same importance at a given point in time. Each needs to be weighted by *w(i, t),* where $0.0 \leq w(i, t) \leq 1.0$, for the *i*-th variable at time *t.* Finally, the overall control strategy seeks to minimize the deviations between observed and desired outcomes and does this by generating the appropriate control signals to each individual attentional mechanism. This global objective is to seek the set of control signals that satisfies:

$$\min_{\Gamma} \sum_{i} w(i,t)\, \Delta_i(K_r^i(t) - K_\gamma^i(t))$$

The "goals" in Figure 3 provide the information on context, task and knowledge needed to dynamically set $K_r^i(t)$, whereas the "priorities" provide information for *w(i, t).* See the figure caption for further detail.

Figure 3 depicts a 2-level control regimen, but there is nothing to limit this to two levels only. It is easily extended to more levels to create a control hierarchy. Moreover, this need not be a strict hierarchy but can take a lattice form if needed. For example, STAR requires both an attention and a task controller [14]. This could be a third level in an extension of the figure. Since the task controller would also oversee memory structures, motor structures and other sensory processing in addition to vision, a branching of control loops would arise. Further development of this is future work but this shows how one might approach this problem.

The control schemes just described represent first steps towards an overall attentional executive for vision. They are more than speculative proposals but still far from being fully proved. There is clearly much to be done before all of the attentional mechanisms listed at the beginning of this paper can be orchestrated to provide observed human behavior. However, we now provide a small look into what the goal might be.

Figure 4 depicts an example of how the various control signals required for some visual tasks might appear. The only part of each signal that is shown is its "on" and "off" point, not magnitudes or any other detail, which have clear importance and need to be properly specified. The signals themselves appear in the middle portion of the figure, with the x-axis representing time. The way to parse this figure, in addition to reading the caption, follows. The time axis shown at the bottom applies to all levels of the figure. Part A is intended to show that control signals depend on the task of the moment. Each of the tasks considered are grouped by what combination of processing cycles on the visual hierarchy (shown in Part C) are needed for their solution (as is more fully described in [1,33]). The color code matches the signals in the middle portion of the figure. The full temporal extent of the visual search task, for example, is the entire red line, while for the discrimination task it is the magenta line. The dashed portion of the task lines represent the pre-stimulus-onset period (where cueing might take place) and the solid part is the period when the stimulus is seen. The various components of Part B in the figure – priming, disengage, attention, etc. – are the specific signals that a control system must generate, with "on" and "off" shown at the appropriate times. These are strictly representative only. Part D shows a series of directed loops. Suppose you are conducting an experiment, say for an identification task. A subject would be given an instruction, then an image. The visual hierarchy would be utilized for cueing, then feedforward processing, then partial feedback processing, and then the trial ends and the sequence repeats. The sequence of trials could be represented by these loops, and the loop with extent from 0 to 215ms would be the loop of sub-processes that together form what the identification task requires. A similar explanation applies for the other loops. Experimental justification for such a task decomposition is seen in [33-38] where observations showing differing tasks leading to different completion durations are documented. There is no use here for open-loop control even though some signals in the figure might seem to imply it. The reason is that due to noise the actual time when one action ends is uncertain and thus setting the start of the next action without monitoring the end point of the previous would not lead to coordinated actions.

Let us return to the example mentioned earlier, that of detecting a parent's face in a single image. In the case of your mother's face in a portrait-like image, it would be an instance of a detection task in Figure 4 involving the first 150ms of the timeline of processing shown. In the case of finding your father's face in a crowd scene this would be an instance of the visual search task which involves the whole time line shown and perhaps several repetitions of it. The structure proposed here is claimed to be rich enough to represent a broad variety of visual behaviors [1,33].

CONCLUSIONS

This paper began with the premise that attentional control is critical for human vision. In contrast to previous

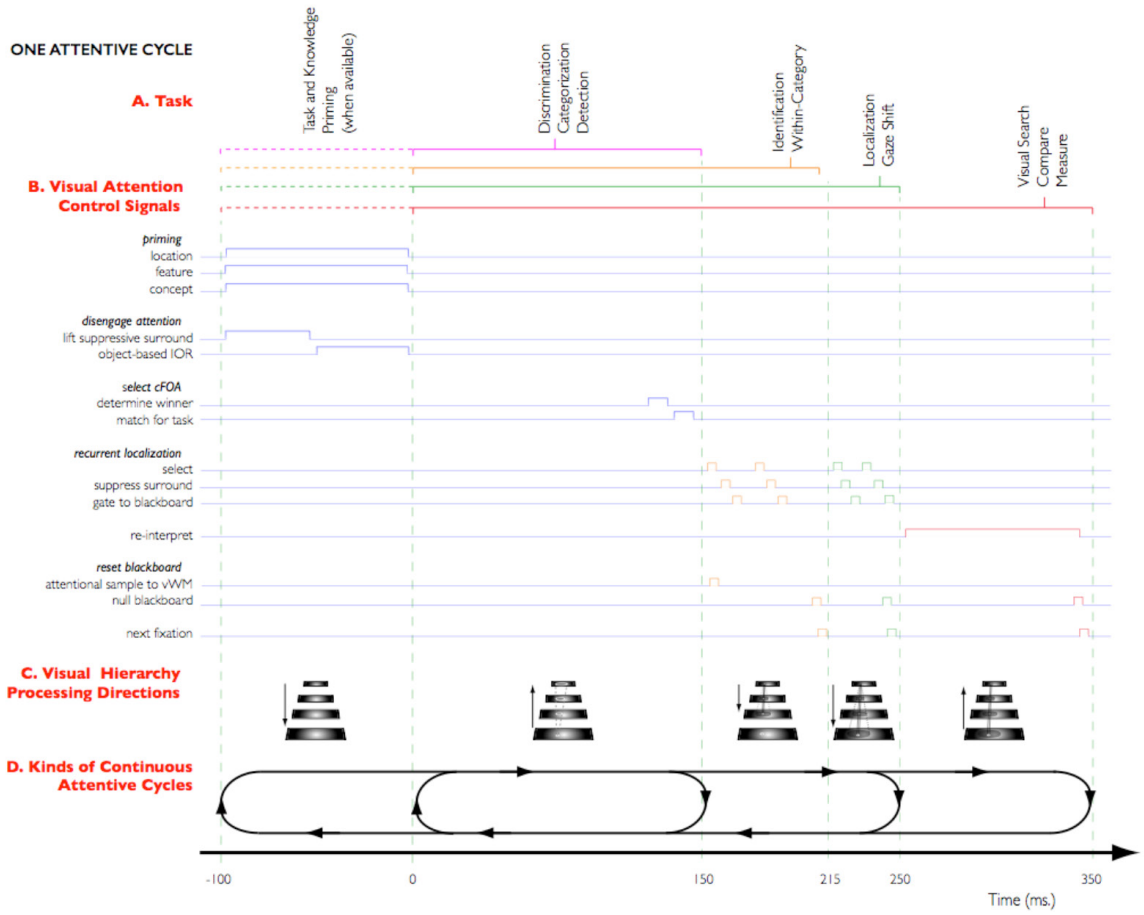

Figure 4. The cyclic control signals required for each of a number of different kinds of visual tasks. **A.** The set of tasks considered appear at the top, with the temporal extent of each shown in a different color. **B.** The set of control signals shown by the blues line to the right of the labels. These blue lines are intended to represent the timing of when the controller must generate an instruction appropriate to the indicated label. For example, in order to select the cFOA (central focus of attention), two actions are required at the end of the first feedforward pass through the visual hierarchy, namely, a selection of the strongest response followed by a matching function to compare what that strongest response represents to the goals of the task. Note how some signals (for example, "null blackboard") have multiple and different colored pulses, one pulse for each layer of the visual hierarchy. **C.** ST has different temporally ordered stages of visual processing [33], beginning with top-down priming due to task instructions or knowledge and context, feedforward processing of visual signal, decision-making to confirm if a task is complete, recurrent processes to reduce network signal interference and permit identification of features or location, search of an image that would entail cycling through various parameterization of the previously listed tasks, and so on. Several experimental results support this conceptualization [1,34]. **D.** These oval circuits are intended to show the different kinds of attention cycles that this sequence of processing stages is capable of representing. Namely, a single attentive processing cycle may or may not include a priming stage, it would end at 150ms if the task was a simple categorization (no attention is needed), or it may require precise localization, etc. This degree of flexibility is a unique feature of STAR.

proposals for an attention executive, the one presented is not based on brain connectivity, brain areas or other particular behaviors. Instead, its foundation begins with the computational characteristics of the structures involved in vision. Based on these, we define several different control goals, each for a specific attentional mechanism. To tie these together, the executive needs a directive, and as shown in Figure 4, the visual task of the moment plays the role of that directive via its representation as a CP. The CP determines how the visual hierarchy is tuned and deployed, how long the task takes to complete, what subtasks are required for its completion, and their ordering.

A key assumption here is that the start and end of each of the behavior elements listed for our toy example

of detecting a parent's face requires a separate and unique signal generated by an executive controller. In other words, none of those behaviors are initiated nor terminated as some kind of emergent process. The reasoning here is that any process from which behaviors might emerge would not also be able to track that behavior's performance, to detect when it fails, nor to then reason about how to repair or refine the behavior so that it can succeed. This follows directly from the nature of emergence: an emergent behavior is the result of some set of sub-behaviors whose sum of effects (and side-effects) provides the desired behavior. Thus, there is no element of the overall system that would know in advance what the behavior should be as would be necessary for solving any specific task, nor would it know how to adjust those elements if a particular arrangement fails to produce the desired behavior (see [39] for more on emergent behavior). In an emergent system, there is little or no central control whatsoever; neurons and brain areas operate in an uncoordinated, asynchronous manner. One would still need to understand how the precision observed in human behavior comes about. A more deterministic process seems required.

Many have recently written about attentional control. A notable starting point is the work of VanRullen and colleagues [40]. They observed multiple periodic functions in human attentional and perceptual behavior and note how these might be orchestrated is an open question. The concept of an attentional cycle is apparent in our proposal. To this point, Pezzulo and Cisek [41] look to control theory as a source of inspiration. They argue for a hierarchy of control loops, using affordances and intentions as basic elements. Our proposal also involves a hierarchy of control loops but the control elements are at a much finer level of abstraction and are more specific to control of attention; but there is consistency in the comparison. Mansouri *et al*. [42] examine which brain areas might be responsible for executive control, motivated largely by the Botvinick *et al*. [43] conflict-monitoring model. Part of our motivation is similar in that the minimization of interference within the visual pathways is one of the control goals. Ours is more specific than the more generic view of cross-pathway interference described by the authors. The authors describe brain areas that would operate to detect conflict; although it is reasonable that such a mechanism might exist, it is different from the interference we try to minimize, which is a necessary consequence of the anatomy of the visual pathways. We do not at this point connect our proposal to specific brain areas because we desire to first demonstrate that our strategy replicates observed attentional behavior. The main difference with each of these other works, and in fact with all others, is our mechanistic dimension, possible because it is developed with Selective Tuning as a foundation. Much empirical verification is required to be sure, but at least here this is possible, whereas with the other models it is not.

Of course, there are open problems with the overall solution we have sketched. What are all the control variables? We have presented three possible ones, but these are neither the full set nor are we sure they reflect single processes in the brain. We have not specified any of the weighting factors nor how they might be determined. Finally, it is uncertain what the best way to represent system goals and knowledge in such a control regimen. It is likely that specific learning paradigms would be relevant here, and useful directions are highlighted in Kumaran *et al*. [44]. Much remains for future work.

The messy reality of attention is that it has many facets and their orchestration is complicated. It might be that with better understanding of how humans attend and behave, the story will become neater. What has been presented here can at least present new hypotheses to test whose conclusions might help tidy things up.

Acknowledgments: The author thanks Thilo Womelsdorf for discussion and for contributions to the early details of the concept, some of which figure prominently here.

Funding Sources: This research was supported by several sources for which the author is grateful: Air Force Office of Scientific Research (FA9550-18-1-0054), the Canada Research Chairs Program (950-231659), and the Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05352).

REFERENCES

1. Tsotsos JK. A Computational Perspective on Visual Attention. Cambridge: MIT Press; 2011.
2. Tsotsos JK. Complexity Level Analysis Revisited: What Can 30 Years of Hindsight Tell Us about How the Brain Might Represent Visual Information? Front Psychol. 2017;8:1216.
3. Kahneman D. Remarks on attention control. Acta Psychol (Amst). 1970;33:118–31.
4. Allport A. Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience 1993:183-218.
5. Egeth HE, Yantis S. Visual attention: control, representation, and time course. Annu Rev Psychol. 1997;48:269–97.
6. Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci. 2002;3:201–15.
7. Rossi AF, Pessoa L, Desimone R, Ungerleider LG. The prefrontal cortex and the executive control of attention. Exp Brain Res. 2009;192(3):489–97.
8. Miller EK, Buschman TJ. Cortical circuits for the control of attention. Curr Opin Neurobiol. 2013;23(2):216–22.
9. Brown JW. The tale of the neuroscientists and the computer: why mechanistic theory matters. Front Neurosci.

2014;8:349.
10. Kotseruba I, Tsotsos JK. 40 years of cognitive architectures: core cognitive abilities and practical applications. Artif Intell Rev. 2018:1–78.
11. Itti L, Rees G, Tsotsos JK, editors. Neurobiology of attention. Amsterdam: Elsevier Press; 2005.
12. Carrasco M. Visual attention: the past 25 years. Vision Res. 2011;51(13):1484–525.
13. Nobre K. Kastner S. (eds.). The Oxford handbook of attention. Oxford University Press; 2014.
14. Tsotsos JK, Kruijne W. Cognitive programs: software for attention's executive. Front Psychol. 2014;5:1260.
15. Ullman S. Visual routines. Cognition. 1984;18:97–159.
16. Rosenfeld A, Biparva M, Tsotsos JK. Priming Neural Networks. CVPR Workshop on Mutual Benefits of Biological and Computer Vision 2018, Salt Lake City UT.
17. Biparva M, Tsotsos JK. STNet: Selective Tuning of Convolutional Networks for Object Localization. ICCV Workshop on Mutual Benefits of Cognitive and Computer Vision 2017, Venice.
18. Wloka C, Kotseruba I, Tsotsos JK. Active Fixation Control to Predict Saccade Sequences. CVPR 2018, Salt Lake City UT.
19. Marr D. Vision: A computational investigation into the human representation and processing of visual information. New York: Henry Holt and Co.; 1982.
20. DiCarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? Neuron. 2012;73(3):415–34.
21. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.
22. Tsotsos JK. "Complexity Level" Analysis of Vision. In M. Brady and A. Rosenfeld (eds.), Proceedings of the 1st International Conference on Computer Vision 1987:46 – 55. London, UK.
23. Tsotsos JK. The Complexity of Perceptual Search Tasks. In N. Sridharan (ed.), Proc. 11th International Joint Conference on Artificial Intelligence 1989:1571 – 1577. August 20 – 25, Detroit, MI.
24. Uhr L. Layered " recognition cone " networks that preprocess, classify and describe. IEEE Trans Comput. 1972;C-21(7):758–68.
25. Jolion JM, Rosenfeld A. A pyramid framework for early vision. Dordrecht: Kluwer; 1994.
26. Rosenfeld A, Zemel R, Tsotsos JK. The elephant in the room. arXiv:1808.03305. 2018.
27. Vorobyov SA. Adaptive and robust beamforming. Academic Press Library in Signal Processing, 3. Array and Statistical Signal Processing; 2014. pp. 503–52.
28. Reynolds J, Desimone R. The role of neural mechanisms of attention in solving the binding problem. Neuron. 1999;24:19–29.
29. Culhane S, Tsotsos JK. An Attentional Prototype for Early Vision. In G. Sandini (ed.), Computer Vision — ECCV'92. Second European Conference on Computer Vision 1992; Santa Margherita Ligure, Italy, May 19 – 22.
30. Tsotsos JK, Culhane S, Wai W, Lai Y, Davis N, Nuflo F. Modeling visual attention via selective tuning. Artificial Intelligence 1995;78(1-2):07-547.
31. Tsotsos JK, Kotserub I, Wloka C. A focus on selection for fixation. J Eye Mov Res. 2016;9(5).
32. Vroemen B, de Jagerl B. Multiobjective control: An overview. Proceedings of the 36th Conference on Decision & Control 1977, San Diego, California USA.
33. Tsotsos JK, Rodriguez-Sanchez A, Rothenstein A, Simine E. Different Binding Strategies for the Different Stages of Visual Recognition. Brain Res. 2008;1225:119–32.
34. Tsotsos J, Womelsdorf T. Visual tasks lead to unique sequences of cyclic attentional signals. Journal of Vision 2016;16(12), 616-616. Available from F1000Research 2016, 5:2467 https://f1000research.com/posters/5-2467
35. Fiebelkorn IC, Saalmann YB, Kastner S. Rhythmic Sampling within and between Objects despite Sustained Attention at a Cued Location. Curr Biol. 2013;23:2553–8.
36. Dugue L, Marque P, VanRullen R. Theta oscillations modulate attentional search performance periodically. J Cogn Neurosci. 2015;27:945–58.
37. Landau AN, Fries P. Attention samples stimuli rhythmically. Curr Biol. 2012;22:1000–4.
38. Desrochers TM, Burk DC, Badre D, Sheinberg DL. The Monitoring and Control of Task Sequences in Human and Non-Human Primates. Front Syst Neurosci. 2016;9:185.
39. Steels L. Components of expertise. AI Mag. 1990;11(2):28–49.
40. VanRullen R. Perceptual cycles. Trends Cogn Sci. 2016;20(10):723–35.
41. Pezzulo G, Cisek P. Navigating the affordance landscape: feedback control as a process model of behavior and cognition. Trends Cogn Sci. 2016;20(6):414–24.
42. Mansouri FA, Egner T, Buckley MJ. Monitoring demands for executive control: shared functions between human and nonhuman primates. Trends Neurosci. 2017;40(1):15–27.
43. Botvinick MM, Cohen JD, Carter CS. Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci. 2004;8(12):539–46.
44. Kumaran D, Hassabis D, McClelland JL. What learning systems do intelligent agents need? Complementary learning systems theory updated. Trends Cogn Sci. 2016;20(7):512–34.