THE EMBODIMENT OF ATTENTION IN THE PERCEPTION-ACTION LOOP

Michael J. Spivey and Stephanie Huette

Introduction

Decades ago, the sciences of mind were busy drawing insights from computer engineering. Cognitive scientists reasoned that if a computer can process information in an intelligent fashion, perhaps humans are exhibiting their intelligence via similar mechanisms. It was thought that we could “reverse engineer” the human mind by drawing an analogy to how a computer is engineered. Of course, this was not the first time that a new and exciting piece of technology had been used as a metaphor for the mind. The Greeks likened the mind to a water pump, eighteenth-century Western philosophers likened the mind to a clock, and then theories of cognition were inspired by the steam engine, then by the telegraph, then by relay circuits, and now the computer (see Daugman, 1993). After using the computer metaphor for several decades now, is it possible that the insights it can provide have all been plumbed?

This chapter describes a series of examples of where the computer metaphor of the mind breaks down, with a special emphasis on attention, and points to an embodied and situated account of mind that can naturally accommodate those problematic phenomena that undermine the computer metaphor. Rather than proposing the next new metaphor for the mind, we instead encourage drawing eclectic inspiration from embodied cognition (to appreciate how the body itself performs some cognitive operations), ecological psychology (to appreciate how the relation between organism and environment produces cognition), dynamical systems theory (to carry out analog information-processing simulations), and cognitive neuroscience (to stay grounded with the real physical material that lies at the hub of cognitive phenomena).

Stage-based models of attention

Early models of attention were based on the assumption that the vast onslaught of raw uninterpreted sensory input that barrages our sensors must be filtered down to a small subset of processed cognitive items that would be analyzed by a central executive with limited capacity. This perspective inspired “filter” models of attention, that immediately drew a distinction between “pre-attentive” filters that processed a full sensory array of information in parallel, and “attentive” filters that processed a condensed set of information serially, one item at a time.
An example of a pre-attentive filter could be a topographically arranged collection of color-sensitive neurons in visual cortex, functioning like a feature-map that detects, say, the color red anywhere in the visual field. Due to its parallel processing, this feature map will locate a solitary red target in the visual field essentially immediately, no matter how many non-red distractor objects are there. Therefore, reaction times do not increase as one adds non-red distractor objects. By contrast, if the target can only be identified by a conjunction of two features, say red and vertical (because there are some red non-vertical distractors and also some non-red vertical distractors), then the “redness” feature map and the “verticalness” feature map will each have multiple regions of activation on them – making pre-attentive identification of the lone target impossible. In this circumstance, the feature integration theory (Treisman and Gelade, 1980) of visual search proposed that attention must combine those feature maps into a “master map,” and this master map is searched serially, as though by a spotlight that focuses on each object one at a time until the target is found. The feature integration theory of visual search was a powerfully influential filter-based account of attention for more than a decade.

However, a number of findings posed problems for the model. For example, some conjunctions of certain types of features appear to elicit parallel search processing, rather than serial (Nakayama and Silverman, 1986). Continuous gradations in similarity between a target and its distractors account for reaction time patterns better than treating visual features as discretely belonging to one feature map or another (Duncan and Humphreys, 1989). Analysis of a million trials across dozens of visual search experiments reveals no evidence for one population of serial search phenomena and a separate population of parallel search phenomena (Wolfe, 1998). And when the visual display spends its first 100 milliseconds configured with a single feature distinguishing the target, before then converting to a conjunction of features, one sees neither a parallel search nor a serial search, but something clearly in-between (Olds, Cowan, and Jolicoeur, 2000). These types of findings all led to treating visual search as belonging on a continuum of efficiency, rather than being implemented by either a pre-attentive parallel filtering mechanism or an attentive serial filtering mechanism. As the passive feedforward filter approach to visual attention failed to account for new findings, it made room for a more proactive goal-oriented approach to attention (Allport, 1989). Rather than the emphasis being on how perceptual systems convert stimuli into interpretations, a new emphasis was building on how attention results from motor systems making real-time demands on what sensory input is needed or expected next (Hommel, Müseler, Aschersleben, and Prinz, 2001; Jordan, 1999; Kawato, 1999; van der Heijden, 1996; see also Gibson, 1979).

**Interactive models of attention**

This more proactive approach to attention raises the question: Where in this goal-oriented account of attention do the goals come from? Desimone and Duncan (1995) proposed a neurally inspired account of visual attention in which neural representations of objects (i.e. Hebbian cell assemblies, or population codes) compete against one another in parallel for the privilege of eliciting their associated motor output, such as an eye movement or a reaching movement. Neural feedback projections from frontal brain regions to visual brain regions can then bias this competition process based on goals and other information from other sensory sources. This “biased competition” account of attention has been extremely influential, and naturally makes a number of empirical predictions that have indeed been reported in the experimental literature.

For example, when a visual stimulus is especially relevant for the goals of a visuomotor task, single-cell recording in monkeys shows that neurons responsive to that stimulus exhibit a higher
firing rate than when that same visual stimulus is presented to that receptive field but is irrelevant to the task (Moran and Desimone, 1985; Motter, 1993). Moreover, fMRI results with humans show that when a tactile precue is spatially congruent with a visual stimulus, activation in visual cortex for that visual stimulus is greater than when the tactile precue is spatially incongruent (Macaluso, Frith, and Driver, 2000). Visual input of a face can trick the process of speech recognition into perceiving a spoken syllable one way or another (McGurk and MacDonald, 1976). Auditory input can trick the visual system into perceiving a visual event one way or another (Sekuler, Sekuler, and Lau, 1997; Shams, Kamitani, and Shimojo, 2000). And spoken linguistic input can make a conjunction search process function as though it were a parallel process, uninfluenced by the number of distractors (Chiu and Spivey, 2011; Spivey, Tyler, Eberhard, and Tanenhaus, 2001). In sharp contrast to Fodor’s (1983) proposal of sensory systems as informationally encapsulated modules, these kinds of findings support a radically interactive and distributed account of attention, where even purportedly “unimodal” perceptual systems are being biased by other modalities (Driver and Spence, 2000; Lupyan and Spivey, 2008; O’Reilly, Wyatt, Herd, Mingus, and Jilk, 2013).

Cascaded models of attention

Even if every subsystem is influencing practically every other subsystem, this leaves open the question of how immediately those influences are transmitted. Traditional stage-based accounts of processing function by having each stage complete its operations before then passing an output to the next stage. This is quite different from a type of system that functions on partial, fuzzy, incomplete, and distributed information that is constantly in flux. It may feel counter-intuitive to think of attention as something this diffuse and partial – rather than it being narrow and directed – but many studies have shown that we operate as best we can with a constant flow of partial bits and pieces of information (Coles, Gratton, Bashore, Eriksen, and Donchin, 1985; for review, see Spivey, 2007).

The cascade model put forth by McClelland (1979) posits a few components to this kind of account, namely that the inputs to a particular level of processing are based on the outputs from the neighboring level(s), and that at each level of processing, the output is continuously available to the neighboring level(s). That is to say, the output is not a singular symbol or conclusion, but rather an unfolding of activation that is continuously being streamed to the neighboring level(s). Thus, the activation at any given level is determined by the connections between levels, as well as the pattern of activation in the neighboring level(s).

A cascaded flow of information from perceptual systems to motor systems is particularly well demonstrated by the work of Gold and Shadlen (2000). While briefly presenting a somewhat ambiguous visual stimulus to a monkey, to which it was trained to respond with an eye movement to one or another response location, they microstimulated cells in its frontal eye fields (FEF) to elicit a neutral direction of eye movement. With more and more time to process the visual stimulus, this electrically evoked saccadic eye movement showed more and more partial influence of the monkey’s voluntary response eye movement, averaged into the angle of the neutral evoked saccade. Thus, over the course of a few hundred milliseconds, the continuous accrual of perceptual information in visual brain areas was constantly cascading into oculomotor brain areas to contribute to the gradual development of an eye movement command. When microstimulation then triggered the evoked saccade in a neutral direction, some portion of that still-developing voluntary response was also active in FEF. The resulting movement was a weighted average of the evoked neutral saccade and the partially accrued command for the voluntary response saccade. Clearly, the motor system is not patiently waiting
to be delivered a completed command from the cognitive system, but is instead continuously participating in the development of the understanding of what a stimulus means for the organism – thus blurring the line between what is “motor” and what is “cognitive.”

This kind of partial processing can be found in many decision-making tasks, where information can cascade through the system even without subjective awareness. Van Rullen and Thorpe (2001) trained participants to rapidly identify whether an animal or means of transportation was present in a picture, displayed for a mere 20 milliseconds on the screen and then masked. Animal and transportation trials occurred in separate blocks, and responses were made via a go, no-go paradigm. Participants were extremely accurate in this task (94 percent correct on average) in identifying whether these objects were present or not. An initial interpretation of this result might be that in a mere 20 milliseconds of input, the feedforward sweep of sensory information appears to be enough to carry out visual object identification in a seemingly context-free manner (before frontal brain regions would have time to respond and send feedback to visual areas). However, this neglects the fact that the task of identifying a pre-specified object’s presence or absence is itself a goal-oriented task. Goals are driven by frontal and prefrontal cortex, which has feedback connections to visual cortex. What is likely happening in this scenario then is not a context-free feedforward sweep whereby some particular category is rapidly identified, but rather the goal of looking for this target, even before a stimulus is presented, involves feedback signals from frontal regions tuning those visual receptive fields to temporarily become “car detectors” or “animal detectors,” focusing processing on features that are associated with those categories.

What then directs where attention goes, if it works in this radically distributed interactive fashion? If we have a system that cascades its information from subsystem to subsystem so promiscuously that we lose track of which signals are feedforward and which are feedback, can we still have something coherent enough to be called “attention”? Rather than conceiving of attention as a “spotlight” directed by a central executive, a more scientifically sound approach may be to conceive of attention as the emergent result of biased competition among multiple partially active representations (Desimone and Duncan, 1995). Rather than assuming that the focus of mental activity at any one point in time somehow derives from one source, like the central processing unit of a computer, perhaps it instead derives from myriad sources. Some of these sources are goals and biases from neighboring neural subsystems. Some of them are biases from immediately perceivable environmental constraints. Some of them are biases from signals delivered by other people in the environment. At longer timescales, some of these biases are social norms and cultural constraints. At even longer timescales, some of these sources are evolutionary influences. The density of perceptual acuity at a particular moment in time — the sense of “who we are” and “what we’re about” at that moment — is something that emerges non-linearly out of the complex interactions among innumerable informational biases competing and cooperating in real time (Spivey, 2013).

Cascaded interaction spreads

With different sensory, cognitive, and motor systems continuously biasing one another in real time, there is no point in time during which a given subsystem is processing its input in a context-free manner. There will always have been some form of contextual bias that was busy influencing the given subsystem immediately before some new afferent sensory input enters it. If continuous distributed interactions between sensory systems make vision no longer an encapsulated module, and audition no longer an encapsulated module, then continuous distributed interactions among cognitive processes and motor processes make cognition no longer a module that is separate from
action – hence, embodied cognition. In fact, given the millisecond timescale at which various motor movements continuously update the sensory input, which continuously updates cognitive processes, which continuously update the ongoing motor commands, perhaps those external actions themselves may not even be informationally encapsulated from cognition. As Sue Hurley (1998, p. 3) put it: “if internal relations can qualify as [representational] vehicles, why not external relations? Given a continuous complex dynamic system of reciprocal causual relations between organism and environment, what in principle stops the spread? The idea that [representational] vehicles might go external takes the notion of distributed processing to its logical extreme.”

A famous example of cognitive processes “going external” comes from Kirsh and Maglio’s (1994) Tetris experiments. They found that novice Tetris players tended to perform something equivalent to “mental rotation” (à la Shepard and Metzler, 1971) of the shapes in the game to determine where to place them. By contrast, expert Tetris players offloaded that cognitive operation onto their environment by using the fast rotate-button frequently. This allowed the computer interface in front of a participant to perform the cognitive operation of image rotation and the neural system in their skull to perform the cognitive operation of perceptual matching. The result was that these experts were significantly faster and more accurate than the novices.

Skeletal motor movements are a useful way to perform cognitive operations, such as pressing keys on an interactive computer interface. But there is a far more ubiquitous motor movement, happening about three times per second, that performs cognitive operations during your entire waking life, no matter where you are, or what you are doing: eye movements. Ballard, Hayhoe, Pook, and Rao (1997) demonstrated that a given eye fixation can function like a “pointer” (in a content-addressable computer memory). Rather than forming neural representations of complex visual objects in memory while viewing a scene, people often simply maintain spatial locations in memory, and send the eyes to those locations to access the “content” in those “addresses.” Essentially, people exhibit a strong tendency to use one’s environment as an external memory store (see also, O’Regan, 1992).

Thus, on the timescale of a few hundred milliseconds, eye movements are changing how the environment impacts your sensors, which changes how you cognize your environment, which changes the next eye movement, which changes the next sensory input, and so on. This is a recurrent causal loop that cannot be unraveled to find the chicken versus the egg. Half of the data transformations that constitute an organism’s attention are actually happening outside the body, in how the body’s movements have altered how light projects onto the retinas. So much so that one could be tempted to suggest that much of your “visual experience” is itself happening in the space around you, rather than solely inside your brain (O’Regan and Noë, 2001; see also Clark and Chalmers, 1998).

**Eye movements and the perception-action loop**

It has long been understood that a there is a close link between visual attention and eye movements. Even when an experimental participant fixates a central dot on a computer screen and covertly directs attention to the periphery for several hundred milliseconds – as is often done in cognitive psychology experiments (Posner, Snyder and Davidson, 1980) – oculomotor systems of the brain are essentially programming an eye movement and then holding back from executing it, and this process is part and parcel of visual attention (Corbetta, 1998; Deubel and Schneider, 1996). In fact, neurons in parietal cortex receive feedback from oculomotor regions signaling an upcoming saccade, such that they are able to shift their receptive fields to the new upcoming region in the visual field and begin responding to a visual stimulus that is not yet in their classical receptive field (Colby, Duhamel, and Goldberg, 1996). This prospective responsiveness to visual
input, in anticipation of a saccade, clearly blurs the line between the motor programming of a saccade planning and the process of covert visual attention. As a result of this close link between these two processes, the vast majority of covert shifts of attention are immediately followed by an eye movement to that same location. Thus, eye movements are quite literally a physical embodiment of visual attention. They allow even a subtle cognitive bias about what is goal-relevant in the environment to be converted – in just a couple hundred milliseconds – into a powerful perceptual bias for what is dominating visual processing, via the high-resolution fovea.

In this way, not only are eye movements highly useful for the experimenter, as a record of what is drawing attention at any moment, but they are also highly useful for the organism itself. A different pattern of eye movements on the same scene will not only reveal a different thought process about that environment (Yarbus, 1967), it can cause a different thought process about that environment. For example, just as recording eye movements during mechanical problem-solving has provided a deeper understanding of mechanical reasoning (Hegarty and Just, 1993; Rozenblit, Spivey, and Wojslawowicz, 2002), it has also provided inspiration for improvement of training with technical devices (Hegarty, 2005).

Even high-level conceptual insights can be jump-started, or missed entirely, by different eye movement patterns on a diagram associated with an insight problem. People who are 30 seconds away from solving a diagram-based version of Karl Duncker’s famous tumor-and-lasers radiation problem tend to make a characteristic pattern of eye movements on the diagram (Grant and Spivey, 2003). Those who are 30 seconds away from giving up on the problem exhibit less of that eye-movement pattern. Interestingly, this pattern of eye movements was not merely an indicator to the experimenters that a person was about to solve the problem, but that pattern of eye movements was actually assisting the person in arriving at the solution. In their second experiment, Grant and Spivey subtly animated the diagram in a way that unconsciously induced that pattern of eye movements, and it doubled the proportion of people solving the problem. Thomas and Lleras (2007) followed up this work with a version of the task where participants were explicitly instructed to move their eyes in a pattern across the diagram. Participants who moved their eyes in the pattern that produced converging lines on the tumor exhibited a higher rate of finding the solution. As one might expect, based on the overlap between attentional mechanisms and eye-movement mechanisms, even when participants don’t move their eyes at all, but instead move their attention covertly in this converging-lines pattern, performance on this insight problem is improved (Thomas and Lleras, 2009). Based on results like these, it may be useful to treat the relation between organism and environment as not only the place where perception and action take place (e.g. Gibson, 1979), but also the place where cognition takes place.

The embodiment of attention

Cognitive science has a long history of drawing inspiration from the computer metaphor of the mind and placing the bulk of its emphasis on how external stimuli influence mental processing, but this is only one half of the story. The other half, of equal importance, is how mental processing influences external stimuli. When attention gets physically embodied in motor movements (even subtle and brief ones like eye movements and other muscle twitches), those movements change the way the environment impacts our sensors in ways that abide by continuous physical laws (Gibson, 1979; Turvey, 1992; Chemero, 2009). The result is that while the environment is catalyzing events in the brain, the brain is simultaneously catalyzing events (i.e. motor movements) in the environment. This autocatalytic loop, whereby the brain and the environment produce an emergent phenomenon called cognition, steers our understanding of
the mind away from the formal logical computer metaphor, where genuine autocatalysis is not logically possible. And it is radically changing how we conceptualize joint action when two people are sharing an environment, and intermingling their perception-action loops (Dale, Fusaroli, Duran, and Richardson, 2013; Spivey, 2012). If we acknowledge that attention in particular, and the mind in general, is an emergent property of an embodied brain interacting with its environment, then the future of cognitive science will require a complex dynamical systems approach to cognition (Spivey, 2007), drawing insights from embodied cognition (Barsalou, 1999), ecological psychology (Turvey and Carello, 1995), complexity theory (Van Orden, Holden, and Turvey, 2003), computational neuroscience (Kello, 2013), and robotics (Pezzulo et al., 2011).

References

Attention in the perception-action loop


