Being in the Workspace, from a Neural Point of View

Comments on Peter Carruthers, "On Central Cognition"

Wayne Wu
Center for the Neural Basis of Cognition
Carnegie Mellon University

In his rich and provocative paper, Peter Carruthers announces two related theses: (a) a positive thesis that "central cognition is sensory based, depending on the *activation and deployment* of sensory images of various sorts" (p. XX, my emphasis) and (b) a negative thesis that the "central mind does not contain any workspace within which goals, decisions, intentions, or non-sensory judgments can be active" (p. XX). These are striking claims suggesting that a natural view about cognition, namely that explicit theoretical reasoning involves direct operations over beliefs, is wrong. Our beliefs, on this natural view, interact with each other to yield new beliefs. If Carruthers is right, beliefs have only an indirect or mediated influence in cognition with sensory states playing the direct role. I think it an important feature of Carruthers' discussion that he draws support from work on the neural circuits and mechanisms that subserve these processes. In what follows, I want to register a few worries about the theses and the empirical evidence adduced to support them.[1]

Let me begin with two terminological clarifications and then highlight a striking feature of Carruthers' view. By "central cognition", Carruthers means cognitive processes

[1] I cannot do justice to the wide array of arguments that Carruthers presents for the theses elsewhere, especially in his 2011 (e.g. evolutionary considerations play a significant role). My discussion will focus on the issues of attention and of working memory raised in his article (this issue).

that deploy *working memory*. We can understand working memory to be a short-term memory store where memoranda are actively maintained to aid ongoing task demands. Such memoranda are conscious at least in the access sense. Second, by "sensory representations" as deployed in cognition, I shall refer to the vehicles of representation, namely the neural realization base called upon in both perceptual experience and perceptual imagination (this is consistent with Carruthers' usage).[2] Finally, note that Carruthers' positive thesis imposes a striking restriction on cognition. The primary materials for cognition are sensory representations. There are, however, two exceptions that Carruthers notes, namely desire and "sensory embedded" judgment. I shall focus on judgment since that is the striking case. Even if concepts tied up with judgments can enter cognition, their influence is always mediated by sensory representations. Thus, they have no direct and independent influence on cognition.

But why should we believe this? The following seems to capture the main argument:

1. Central cognition involves the deployment of the constituents of the global workspace.
2. Only sensory states can enter the global workspace (all other states requiring sensory mediation).
3. Negative Thesis: There is no workspace where propositional attitudes can interact (i.e. they cannot enter the workspace).

[2] There's going to be a bit of sloppiness in what follows between representational content and representational vehicles. When we speak of the property of neurons and neural assemblies, we are speaking of the vehicles. As the empirical matters will loom large, it is worth keeping this in mind.

4. Positive Thesis: Central cognition is sensory based.

Premise 2 is sufficient for the Negative Thesis and premises 1 and 2 entail the positive thesis. I shall argue that there are reasons to doubt premise 2.[3]

The concepts of working memory and the global workspace play a central role in Carruthers discussion so I want to enter a clarification about common spatial metaphors like a *store* or a *stage*. Talk of working memory and the global workspace originated at a psychological, functional level of analysis, but neuroscientists have taken the same ideas on board. Our thinking about working memory owes much to Alan Baddley who characterized working memory not as a unitary storage system but as involving distinct components including a central executive that keeps tabs on sensory slave systems. Here are three recent depictions of working memory/global workspace that show a transition in different levels of analysis. The first is from Baddeley's (2010) updated conception of working memory:

Figures redacted due to copyright.

Consider Carruthers' (2011) depiction of working memory as within global workspace:

Finally, Sergent and Dehaene (2004) have modeled the Global Workspace as follows:

[3] It is worth noting that there are other ways one might arrive at the positive thesis. For example, one could argue that all concepts are sensory. This is not Carruthers' route since he allows for amodal concepts. Alternatively, one might argue from introspection that cognition has a sensory phenomenology to the claim that its materials are sensory. This is also not Carruthers' route, or at least not the primary route. Rather, the issues he adduces are drawn from experimental cognitive science.

One thing you might note in this sequence is the progression from a more abstract, functional architecture (Baddeley), to another that is more committal about participating neural systems (Carruthers) to one that clearly is thinking about neural circuits (Sergent and Dehaene). The other thing to notice is that the last two depictions suggest that there literally is a distinct store or stage that serves as the crucial hub to broadcast information to other systems. This leads naturally to the question: where in the brain is working memory or the global workspace? I want to focus on that idea by considering a current debate about working memory.

On what Postle (2006) calls the "standard model" of working memory, there is a specialized system that serves as a storage buffer. This store was initially located in pre-frontal cortex (PFC) due to the finding of sustained neural activity in this region during delay periods in tasks where the subject has to keep a previously perceived stimulus in memory (e.g. in a *match-to-sample* task where subjects have to match a subsequently presented stimulus to one that was previously presented and thus keep the latter in memory). There is, however, some work that suggests that PFC's function is more executive rather than storage (see Miller and Cohen (2001) for this view).

But should we be looking for a determinately localized store called upon in every working memory task? An alternative, advocated by Postle and by D'Esposito (2007) involves a dynamic conception of the working memory store. Consequently, there is no single unitary store, some concrete localized memory buffer that counts as working memory. Rather, when working memory is deployed, the nature of the task determines the areas of the brain that are activated to maintain information to serve the task. So, if the task is visual, then relevant visual areas will be activated and when they are, they are

part of the working memory store. Similarly, if the task if auditory, then relevant auditory areas will be activated and when they are, they are part of the working memory store. This is to say that the location of working memory storage is dynamic, depending on the task and the neural regions that are needed to perform that task.

The same point can be made in respect of the global workspace. As is clear in both Carruthers' and Dehaene's depiction, there is a postulated stage in the workspace distinct from say perceptual systems. Other systems feed into the workspace which stands as an informational hub from which fed contents are broadcast. But despite the suggestion latent in their depiction, Sergent and Dehaene (2004) articulate something closer to the Postle and D'Esposito view in respect of the global workspace:

> A consequence of this [global workspace] hypothesis is the absence of a sharp anatomical delineation of the workspace system. In time, the contours of the workspace fluctuate as different brain circuits are temporarily mobilized, then demobilized. It would therefore be incorrect to identify the workspace, and therefore consciousness, with a fixed set of brain areas. Rather, many brain areas contain workspace neurons with the appropriate long-distance and widespread connectivity, and at any given time only a fraction of these neurons constitute the mobilized workspace. As discussed below, workspace neurons seem to be particularly dense in prefrontal cortices (PFCs) and anterior cingulate (AC), thus conferring those areas a dominant role. However, we see no need to postulate that any single brain area is systematically activated in all conscious states, regardless of their content. It is the style of activation (dynamic long-distance mobilization),

> rather than its cerebral localization, which characterizes consciousness. This hypothesis therefore departs radically from the notion of a single central `Cartesian theater' in which conscious information is displayed

Consider Jonides et al. (2008): "STM [short term memory, including working memory] engages essentially all cortical areas—including medial temporal lobes—and does so from the earliest moments, though it engages these areas differentially at different functional stages" (215). Indeed, Jonides et al. argue that working memory and long term memory *share the same representations*, the main difference residing in whether the memory is supported by neural activity (working memory) or by strength of synaptic connections (long term memory). We might say that working memory reflects the activity of populations of neurons, something dynamic; long-term memory reflects the structure and properties of their physical connections, something more static.

Let's now reflect a bit on the idea of something's *being in the global workspace.* To get at this, I am going to deploy the idea of *computation* in one of the senses drawn on in computational neuroscience. There is much to be said here, and no space to say it adequately, but it will be enough to have an idea that draws on practice in computational neuroscience. Roughly, we shall focus on a computational model from a neuroscientist who provides the model as a description of the activity of neurons or populations of neurons in terms of the information they carry. These neurons are characterized as computing a defined function (e.g. one that implements edge detection). Thus, neurons are part of the computation in the sense that the information they carry is processed so as to compute the relevant function, at least according to the computational description.

Regions that materially support such computations but do not store and provide information over which computations are performed play a non-computational role. Such non-computational regions may affect the properties of the computation, say increase its speed, but do not provide information for the computation.

It is natural then to speak of information as being in the global workspace. Information is processed by the workspace and subsequently broadcast for other purposes. The notion of information here is ambiguous between (i) a *semantic* notion of information, one that admits of accuracy conditions and has its home in our discussion of the propositional attitudes and (ii) an information-theoretic notion, such as that due to Claude Shannon, where (*mutual*) information is plausibly something that neurons transmit, or one that is closer to Fred Dretske's notion of *indication*. In fact, as information in either sense is abstract we can focus on the vehicles that carry such information as what literally are in the workspace qua neural architecture. Thus, Shannon information can be in the workspace in the sense that there is an electrical signal carried by neurons that are involved in some computation of interest.

We began the discussion, however, by talking about the possibility that propositional attitudes, namely mental *states,* occupy the global workspace. This idea is harder to understand since we don't expect the states to literally move into the workspace, understood as some fixed neural region. Here, the dynamic conception opens up another option. Let us assume that the vehicles of propositional attitude contents are vehicles realized in appropriate areas of the brain. When mental states are *tokened,* there is an activation of their neural vehicles. What we have in such a tokening is an *event* such as the electrical activity of neurons. This neural event is part of the workspace when it is

part of a dynamic network that grows and contracts as needed to perform a given task, *so long as the event plays the right computational role*: to provide information over which computations are performed. For example, the regions in question might be activated to carry information over time that would then be used when appropriate in task computations. Propositional attitudes could in that sense become part of the workspace, under certain task conditions, namely when their vehicles are part of it. On this view, there could be a workspace where propositional attitudes can interact.

It seems to me then that the central issue is not whether propositional attitudes can figure in the workspace, but whether their cognitive role is *always dependent on sensory representations*. This is the main import of the second premise above. What evidence do we have for this claim? In the paper, Carruthers notes some relevant neurobiological evidence: (a) the observation, say via neuroimaging, that sensory areas are co-activated with deployment of working memory and, more importantly I think, (b) the underlying circuitry and functions that serve working memory tasks, with emphasis on the role of attention and its ties to "mid-level" sensory areas.

Why think that attention is tied to working memory? There are empirical arguments for this link, but I want to put forward a conceptual argument based on a conception of attention which I think is congenial to cognitive scientists because it is drawn from experimental practice: attention is a specific kind of selection process, namely selection for *action* or *task*. The idea is that when subjects are asked to perform a behavior in an experiment, certain information is relevant and other is not, and successful performance of the experimental task requires only relevant information be selected. Indeed, when experimentalists work on attention, they draw on a set of experimental

paradigms involving well-defined tasks that are designed so as to create conditions in which attention is deployed in a specific way. Accordingly, for paradigms such as filtering information or visual search, the subject's following the task instructions *suffices* for them to deploy attention in a specific way. In general, these tasks create conditions for selection of task relevant information that suffice for the subject's attending to that information. Notice that this sufficient condition does not come from nowhere, but is derived from well-defined experimental practice. So, selection for task is a form of attentional selection.

Arguably, many tasks will require the deployment of working memory not just for fleeting perceptual inputs but also of items remembered long term that must be recalled and kept actively in mind. Selection of specific long-term memories via recollection can also be understood as a form of attention in the sense of selection for task: the large amount of long-term memoranda is analogous to the large amount of perceptual inputs against which we must be selective. In both cases, to perform the task, only a subset of possible inputs must be selected. Thus, subjects must select task-relevant perceptual and mnemonic information and maintain such information to guide their behavior. In this way, attention is implicated in working memory as a by-product of the selective capacities that must be deployed in any intentional behavior, certainly those studied in the lab.

Carruthers seems to suggest something like the following argument with emphasis on attention:

1. Central cognitive tasks implicate the involvement of working memory.

2. Working memory implicates the involvement of attention.
3. Attentional circuitry in primates has as its central waypoint mid-level sensory areas.
4. So: performance of cognitive tasks implicates activation of mid-level sensory areas.
5. So: performance of cognitive tasks implicates deployment of mid-level sensory representations.

If premises 1-3 are true, then 4 is likely to be true. The circuitry would explain the activation of sensory regions in any working memory task and why those activations are necessary for working memory, for they are necessary for attention. Note, however, that 5 is stronger since it emphasizes *deployment* of sensory representations, and it turns out that "deployment" has two potential meanings in this context (see below). Moreover, 5's truth is not settled simply by seeing a region "light up" in neuroimaging of the subject's brain during a task, since the activation could be epiphenomenal. So 5 is a fact to be discovered (if it is a fact). Finally, 5, even if true, isn't sufficient for the positive thesis:

6. Performance of cognitive tasks is based on sensory representations.

This is due to the ambiguity in "deploys".

Let's accept Carruthers' claim that the circuitry in the primate visual system connects areas controlling attention top-down to mid-level sensory cortical areas. So, given 1-3, we should expect activation of sensory areas in any working memory task.

This allows for top-down attentional effects in perceptual *and* non-perceptual systems, so long as top-down attentional modulation in both cases can be transferred via connections tying regions to mid-level sensory cortical areas. Visual attentional effects have been identified as early as the lateral geniculate nucleus (LGN), the thalamic relay point from the eyes which projects to visual area V1, and Stefan Treue (2003) suggests that "practically all cortical visual information processing is shaped by top-down attentional influences; a purely sensory representation of the visual environment does not appear to exist in primate visual cortex" (428).

Carruthers' holds that there are amodal concepts, and attention must also be able to influence amodal conceptual areas for the same reasons given above, namely due to selection for task. There are conceptual tasks that require the selection of appropriate concepts and conceptual states for task performance, and this selection for task is a form of attention. Given the proposed circuitry that we are assuming for the sake of argument, top down attentional effects on conceptual areas will also be sensorily mediated. In other words, we have a causal chain like this:

top-down attentional signal → sensory areas → conceptual areas.

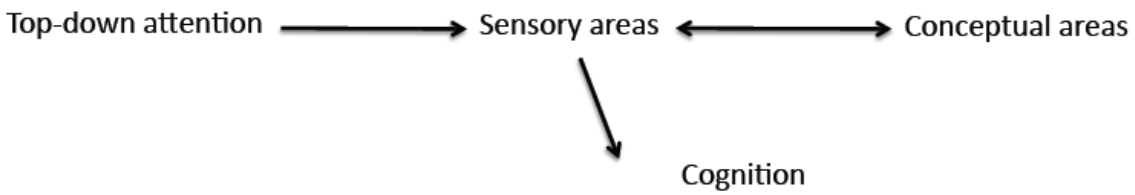
Accordingly, 5 may be true because visual areas, in this case, are deployed to mediate attentional effects to other connected areas. That is to say, any conceptual task will implicate the deployment of sensory areas. "Deployment" here means *used to propagate attentional selectivity* to other relevant areas, but this is not the relevant notion of deployment that the positive thesis requires. Sensory areas are not supposed to be merely

conduits for attentional effects, but are supposed to be the *materials* for central cognition. This means, presumably, that the information that they carry is deployed in the relevant cognitive computations, which will typically involve their being held in working memory. Only on this computational reading of "deployment" would sensory areas provide the foundation for cognition.

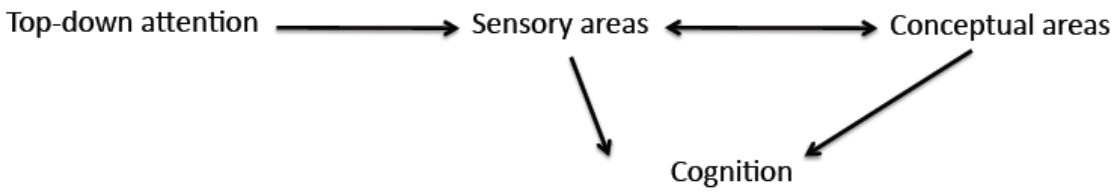
Selectivity of the sort connected to attention, however, is distinct from computation over information.[4] So, for all the imaging shows, it might be that visual areas play only an attentional but not a computational role: they help activate regions that carry computationally relevant information but they themselves don't provide this information over which computations are performed. To see this, recall an early view of attention as a *filter* (Broadbent 1958) allowing only task relevant information to proceed for further information processing. Attention, on this view, does not provide information, it only sifts through it, selecting what is task relevant and blocking off everything else (ideally). If the distinction between mere selectivity and computation is clear, then mere observation of neural activity in visual areas will not settle questions about selectional versus computational role.

Thus, we can agree with Carruthers' that the activation of sensory areas is not epiphenomenal but plays some causal role. Still, this agreement leaves open the two alternatives that we are considering. Thus, on the one hand, we have the model that Carruthers argues for:

[4] That attentional selection is distinct from relevant computational role seems to be why top-down attentional modulation of visual areas by a subject's goals fails to count as a form of the cognitive penetration of vision.



On this model, sensory areas always mediate the role of concepts in (central) cognition (see also the cognitive architecture he depicts above where there is no arrow from conceptual systems to the global workspace; only sensory systems have access to it). On the other hand, there is the model held by his opponent:



The fundamental difference is, pictorially, a single arrow, namely one connecting conceptual areas, the building blocks of the propositional attitudes like beliefs, to cognition. Note also that both models are consistent with co-activation of sensory areas in every working memory task, given the role of attention.[5] Thus, Carruthers' opponents will agree and disagree when he writes:

> attention directed at mid-level sensory areas is a necessary (and perhaps sufficient) condition for global broadcasting to occur, and hence for entry of a given conceptual representation into working memory…If this is so, then sensory

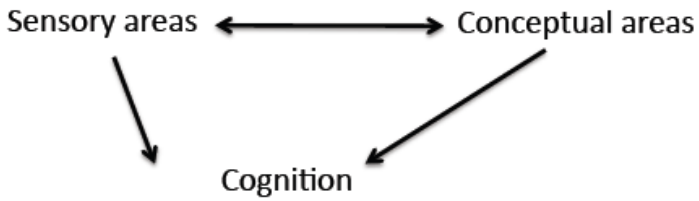
[5] Note that there is a third model where *only* conceptual areas feed into cognition. Perhaps that view is obviously too strong.

> activation is by no means ancillary to the operations of working memory, but rather constitutes its very foundation (XX).

The opponent agrees that sensory activation is not ancillary to the operations of working memory but argues that it does not follow that sensory activation constitutes the very foundation of cognition, or even if it does, it does not follow that it constitutes the *only* foundation. It is consistent with the opponent's model that even if sensory activation is necessary for working memory, propositional attitudes can play a direct role as well. Thus, for all the evidence shows, propositional attitudes can be active in the global workspace and thus, in central cognition.

Carruthers might point out that he has allowed for concepts gaining access to working memory, namely when they are "bound into the content of a perceptual or imagistic state" (XX). This might be found in "sensory embedded" judgments, as when one sees a figure *as a rabbit.* Thus, the access of concepts to working memory is still sensorily mediated. I don't, however, fully understand what "binding" or "embedding" in this context mean (e.g. "conceptual information can be bound into perceptual states and broadcast along with them", XX), and this is not explained. Consider one aspect of the proposed architecture on Carruthers model:

How are we to understand the binding of conceptual contents to sensory contents on this model? Beyond metaphors, what does it mean? If I had to venture a guess on how a psychologist might understand talk of binding, it would depend on the following architecture:



Binding on this model would be *coactivation* of two sorts of representations where their informational contents both continue down in the computational chain into cognition. I hear sounds as having a certain meaning because of the co-activation of sensory, syntactic, and semantic representations that *all* feed forward. There is in this sense binding of conceptual and nonconceptual contents, but it is consistent with the dynamic model above and assumes an architecture that allows conceptual areas direct access to downstream processing relevant to central cognition. If "binding" has some other implementation and meaning, then that should be explained. Otherwise, Carruthers' opponent can claim that only she has adequately explained what binding comes to.

The difference between the two architectures is clear: one allows belief and other propositional attitudes direct access to central cognition; the other does not. So how might we settle the issue? Given the neurobiological considerations that Carruthers' adduces, one relevant form of evidence concerns the role of brain regions that realize amodal conceptual representations. There is some controversy over the existence of such regions and, for those who hold that they exist, where precisely they are localized (e.g.

the anterior pole of the temporal lobe or relevant areas of prefrontal cortex). So there are *existence* and *localization* questions that must be addressed (we assume, with Carruthers, that the answer to the existence question is yes, so we shall ignore it). But say that you settled the localization issue, namely identify the neural seat of amodal conceptual representations. A further question would be whether those areas have the right connectivity with other regions to fulfill the computational role proposed by one of the two models under consideration or whether they lack the connectivity necessary to fill those functional roles. Call this the *connectivity* question. A third question, assuming that one can settle the previous questions is whether the observed activations of any of these regions counts as playing the postulated computational role: the regions serve as a working memory store that aids task demands. Call this the *computational-mechanistic* question. Finally, a fourth question is whether any role for amodal representations we uncover always depends on concurrent sensory representations where both forms of representations play a computational role (ruling out mere spreading activation or attentional effects). Call this the *binding* question.

I do not see that the issue between the two models can be settled without answering all of these questions, nor do I see that the evidence Carruthers' has presented provides a clear answer to these questions, one way or another. Of course, that's not his fault. The answers must be provided by experimental work from cognitive neuroscience. While Carruthers' has presented some suggestive evidence, I think the issue of which model is correct is still quite live, at least from the neural point of view.

There is a general lesson here. I agree with Carruthers on the relevance of empirical work to addressing some central questions about cognition, perception, and the

mind in general. He has done an impressive job in assembling relevant empirical data. My own sense is that in many ways, neuroscience hasn't yet done enough work to give us results that can settle the larger philosophical issues, at least as much as empirical work can contribute to settling them. Or maybe settling is too much to ask, so the point is that neuroscience hasn't given us enough to push us firmly in certain directions. To provide us with evidence that can clearly confront different philosophical views, we need much more detail than neuroscience can currently give us: we need to localize regions that realize relevant representational vehicles; we need to understand the circuitry between various regions; we need to understand what information these vehicles carry; we need to understand the transformations and computations each region performs; we need to understand how the regions interact, the time course and sequence of their interactions, and so on. There are alas scarcely few areas of interest to philosophers where this demand is met, and so there is a way in which science still can't quite help us philosophers on the most significant matters (maybe that's a little too pessimistic; I prefer "sober"). Still, I am sure that getting to the point where the empirical work does shed clearer light on philosophical theses about cognition will require the joint work of philosophers, neuroscientists, and psychologists. That is an exciting project I think worth engaging in.[6]

[6] My thanks to Peter Carruthers for discussing these matters with me and to Todd Ganson for the opportunity to engage Peter's interesting ideas in the very enjoyable forum of the Oberlin Colloquium.

Bibliography

Baddeley, Alan. 2010. "Working Memory." *Current Biology: CB* 20 (4) (February 23): R136–140. doi:10.1016/j.cub.2009.12.014.

Broadbent, Donald Eric. 1958. *Perception and Communication*. Pergamon Press.

D'Esposito, Mark. 2007. "From Cognitive to Neural Models of Working Memory." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (1481) (May 29): 761–772. doi:10.1098/rstb.2007.2086.

Jonides, John, Richard L Lewis, Derek Evan Nee, Cindy A Lustig, Marc G Berman, and Katherine Sledge Moore. 2008. "The Mind and Brain of Short-term Memory." *Annual Review of Psychology* 59: 193–224. doi:10.1146/annurev.psych.59.103006.093615.

Miller, E.K., and J.D. Cohen. 2001. "An Integrative Theory of Prefrontal Cortex Function." *Annual Review of Neuroscience* 24: 167–202.

Postle, B R. 2006. "Working Memory as an Emergent Property of the Mind and Brain." *Neuroscience* 139 (1) (April 28): 23–38. doi:10.1016/j.neuroscience.2005.06.005.

Sergent, Claire, and Stanislas Dehaene. 2004. "Neural Processes Underlying Conscious Perception: Experimental Findings and a Global Neuronal Workspace Framework." *Journal of Physiology, Paris* 98 (4-6) (November): 374–384. doi:10.1016/j.jphysparis.2005.09.006.

Treue, Stefan. 2003. "Visual Attention: The Where, What, How and Why of Saliency." *Current Opinion in Neurobiology* 13 (4) (August): 428–432.

Wu, Wayne. Forthcoming. “Visual Spatial Constancy and Modularity: Does Intention Penetrate Vision?” *Philosophical Studies*: 1–23. doi:10.1007/s11098-012-9971-y.