Review of Philosophy and Psychology (2021) 12:199–225
https://doi.org/10.1007/s13164-020-00507-4

Working memory is not a natural kind and cannot explain central cognition

Javier Gomez-Lavin[1]

Published online: 28 August 2020

Abstract

Working memory is a foundational construct of cognitive psychology, where it is thought to be a capacity that enables us to keep information in mind and to use that information to support goal directed behavior. Philosophers have recently employed working memory to explain central cognitive processes, from consciousness to reasoning. In this paper, I show that working memory cannot meet even a minimal account of natural kindhood, as the functions of maintenance and manipulation of information that tie working memory models and theories together do not have a coherent or univocal realizer in the brain. As such, working memory cannot explain central cognition. Rather, I argue that working memory merely redescribes its target phenomenon, and in doing so it obfuscates relevant distinctions amongst the many ways that brains like ours retain and transform information in the service of cognition. While this project ultimately erodes the explanatory role that working memory has played in our understanding of cognition, it simultaneously prompts us to evaluate the function of natural kinds within cognitive science, and signals the need for a *productive pessimism* to frame our future study of cognitive categories.

1 Preliminaries to the current project

Working memory is thought to be a domain general capacity that enables us to keep information in mind and to use that information in the service of our goals (Baddeley 2007). A classic example of it in action occurs when you try to keep a phone number in mind, where many people will rehearse the digits using internal speech. Working memory is attractive to philosophers precisely because it provides a scientifically vetted construct that is implicated in and that may explain central cognition—that is, *thought*. Proponents of central cognition argue that the mind, in order to flexibly solve problems and reason, must possess a neutral workspace where thoughts from across the mental economy can be brought to bear on one another (Carruthers 2014). Fodor's (1983)

✉ Javier Gomez-Lavin
jgomezlavin@gmail.com

[1] The University of Pennsylvania, Philadelphia, PA, USA

isotropic mental architecture and Evans' (1982) "generality constraint" mandating the inferential "promiscuity" of concepts are some of the most explicit formulations of this account of the mind, which stretches back and features in many prominent theories of cognition and agency.[1] One might even argue that philosophers who make no mention of working memory are committed to something analogous. From neopragmatists like Brandom (2001) and McDowell (1996) who see human thought as playing out amongst an internal space of reasons, to the unified mentality of the Kantian rational agent (Korsgaard 2009), to Aristotle who, in *De Anima,* proposes that thought—in any form—cannot occur without a capacity, *phantasia*, to form and retain images from past experience (431a16; Aristotle and Hamlyn 1993)—working memory promises to be the most likely naturalization of these philosophical commitments.[2]

One might consider the substitution of empirically vetted constructs for undertheorized commitments to be good practice; however, the success of such a substitution relies primarily on whether the relevant contents are preserved in the transition. That is, if our best empirical understanding of working memory does not capture the notion that the mind has a reflective center then theories—including influential first-order accounts of consciousness (Dehaene 2014; Prinz 2012) and reasoning (Evans and Stanovich 2013)— built upon this assumption are compromised. Of course, this will require that we carefully unpack the technical details of working memory to render it amenable to philosophical analysis.

This paper argues that, in contrast to both the prevailing understanding in neuroscience and recent explicit claims made by philosophers, working memory cannot explain central cognitive processes as it merely *redescribes* them. To arrive at this conclusion, I employ the apparatus of *natural kinds* as a conceptual template that helps to render the explanatory deficiencies in contemporary models of working memory explicit. First, I show in section 2 that working memory is treated as a natural kind, that is, as an objective proper part of the mind realized in its neural matter and structure (e.g., Carruthers 2015). In section 3, I introduce Khalidi's property-cluster theory of natural kinds, which presents the most congenial criteria for proponents of working memory's kindhood: it need only capture a set of causal and determinate properties that license productive inferences (Khalidi 2013). If working memory fails to meet the low bar set for kindhood by this account, it will inevitably fail under stricter essentialist scrutiny. Section 4 provides a tour through three prominent models of working memory,

[1] As a cursory gloss, both of these projects aim to identify fundamental properties characteristic of thought, where Fodor's isotropy describes the possibility that the content of any one thought might be relevant to any other, and Evan's constraint details how any number of properties might be predicated to any object of thought. As Carruthers (2014) makes explicit, the possible multiplicity of relations that thoughts and their properties can bear to one other can be captured by the structure of a multimodal mental workspace, which he identifies as working memory (144). Thanks to a reviewer for prompting this clarification.

[2] This is, of course, only a subset of positions in the philosophical literature that plausibly rely on features predicated of working memory. As an anonymous reviewer has pointed out, scholars who use the notion of a "specious presence" or a "circuit of consciousness" may be reliant on something that more closely resembles the modern psychological formulation of working memory (e.g., Ladd 1887; James 1890; Husserl 1964; Bergson 1990), and, reviewing this correspondence is merited in future work. However, by focusing on the authors above, we can showcase the breath of possible theories that make use of a unified, domain-general mental workspace that enables flexible thought. Finally, while it seems odd to suggest that Aristotle employs something like working memory in his description of the phantasia, the connection and striking parallels between the two concepts merits considered exploration, as both are said to retain perceptual information, are crucial to the process of long-term memory formation, and integral to every instance of deliberative thought.

showcasing its gradual fragmentation from a domain-general unitary memory store to an emergent feature of the brain. While this fragmentation may alone serve as sufficient evidence to reject working memory's candidacy as a natural kind, we can also isolate a functional dyad common across these models, which may serve as a basis for a *generic account* of working memory; namely, as the *maintenance* or *manipulation* of information used to guide behavior (Cowan 2017). In Section 5, I demonstrate that the maintenance of task-relevant information occurs throughout the cortex and is realized by many mechanisms, undermining its ability to delineate and ground a domain of inquiry unique to working memory. Instead, the maintenance of information, ubiquitous as it is, serves as the currency of cognition in general. In response to a recent defensive move by cognitive neuroscientists to restrict working memory, properly so-called, only to the *manipulation* of information (Postle 2016), I argue in that any appeal to *manipulation* either devolves into an account of maintenance, with all its concomitant problems, or it circularly appeals to those higher cognitive processes of reasoning and inference that working memory is supposed to explain. In either case, we are left with an explanatorily empty term which at best only offers a terminological variant of *cognition* wholesale, and at worst actively obfuscates our search for the mechanisms and processes that underwrite central cognition. I conclude in section 6 by summarizing the impact of this polemical project, dispelling some misconceptions, and reflecting on the many ways brains can hold and transform information and what this holds for future work on central cognition.

2 Why consider working memory a natural kind?

In the past half century, working memory has transformed from just one of many simplistic models that explained and predicted performance on a series of mnemonic recall tasks to a central feature of our psychology, such that every modern introductory textbook has a section dedicated to it (Gazzaniga et al. 2009). That is, in contemporary psychology and neuroscience working memory is treated as a *natural kind* where its associated properties, and how they are implemented in the brain, are taken to mirror the objective organization of the mind. Of course, researchers' confidence mapping structural and computational properties of the brain to the functional features of working memory does vary. Psychologists often hedge the autonomy of their models from any explicit mention of where and how they are implemented, while encouraging the search for the neural "sites" or correlates (Baddeley and Logie 1999; Baddeley 2010). Neuroscientists tend to be far more zealous in their mappings; however, this metaphysical move from model to implementation didn't occur without due diligence. Neuroscientists have been able to correlate performance deficits in a task to lesions or the stimulation of targeted neural populations, both longstanding tools of contemporary neuroscience (Bogdanov and Schwabe 2017; Postle et al. 2006). The point is simply that—in the empirical domain—the concept of working memory has morphed from a rudimentary boxological model to a functionally distinguishable and causally situated part of the mind with extant homologs in other mammalian, avian, and possibly cephalopod species.[3]

[3] Cf. Carruthers' (chp. 8, 2015) excellent review of avian proception, and Godfrey-Smith's compelling account of cephalopod intelligence (2016).

Philosophers have also identified working memory as a natural kind. Carruthers, in his recent book on the topic—indeed the first thorough philosophical treatment of working memory—conjectures when discussing dual-system accounts of reasoning that, "while System 1 comprises a heterogenous set of processing systems, System 2 has some claim to be considered a natural kind. For it can largely be identified with the working-memory system, which surely qualifies as such" (Carruthers 2015, 180). For Carruthers, working memory and top-down attention are natural kinds as they consist in a stable set of functional and neural components across individuals, with extant homologs across species; and in the case of working memory, it is a *real,* distributed system with attention at its core, or "essence," which creates and maintains multimodal, sensory-contexts in a globally broadcast state through similar attentional "boosting" features present in perception (Carruthers 2015).

Before we evaluate whether working memory makes a good natural kind candidate, it's important to divorce this discussion from the orthogonal question of whether episodic memory, or memory more generally, is a natural kind. There already exists a wide literature on this subject, with most commentators arguing that memory, and episodic memory more specifically, is not a coherent kind concept (Michaelian 2011, 2015; Klein 2015; Cheng and Werning 2016). However, *working memory* is often set aside by these authors, as they generally acknowledge that the functions and realizers of working memory are likely distinct from other forms of memory. The following statement from Michaelian is representative, "I largely disregard working memory, which seems to be a basically distinct phenomenon" (Michaelian 2011, 186). So, is working memory a natural kind? First, we'd do well to explain why we care whether working memory is a kind at all.

3 Property cluster theories of natural kinds

Before we turn to our central question it's important to consider the philosophical elevation of kind terms in general. Why do we care whether working memory is a natural kind? In the first place, and as we noted in the previous section, it is because working memory is *already* assumed to be a natural kind by philosophers and treated as such by neuroscientists. Conveniently, the framework of natural kinds also provides us with a template that we can use to evaluate working memory's explanatory leverage. This reflects a methodological strategy in philosophy of science: natural kinds are supposed to be anchored in the causal structure of the world and are explanatory in virtue of this relationship. At the same time, as there are several accounts of natural kinds on offer, we must come up with a principled way to settle on one view moving forward. This argumentative strategy justifies why I have chosen the most congenial characterization of natural kinds in the offing; namely, Khalidi's property cluster account (Khalidi 2013; Khalidi 2015). This view, which is unpacked below, presents minimal criteria for kindhood, such that if working memory fails to meet these criteria it assuredly will fail under the stricter demands entailed by other views. In turn, I am not suggesting that this property cluster account is the only, or most correct, or in any way the *best* account of natural kinds, it simply presents the easiest set of criteria for a kind candidate to meet. Substituting your preferred theory of natural kinds will yield the same result, where working memory will not meet the threshold of kindhood, except by

pure stipulation. Finally, to be very clear, I am not interested in the epistemic role or metaphysical position of natural kind terms in general, rather my claim is that once we take the psychological and neuroscientific evidence into account, what we call working memory does not map onto a single or even a coherent set of features that we can track in the brain, but instead characterizes what brains do a good majority of the time. As such, providing "working memory" as an explanation of some cognitive phenomenon that we're interested in, such as reasoning or consciousness, amounts to little more than saying that the brain is responsible for it.

Khalidi's property cluster account of natural kinds is a variant of Boyd's (1999) influential, homeostatic property cluster theory. Under this theory, as its name suggests, natural kinds refer to a cluster of causally linked properties, with no one property being "essential" or necessary and sufficient. On the whole, this and related views aim to describe the central processes or mechanisms that cluster a set of relevant properties, allowing them to reach homeostasis.[4] Khalidi's theory carries over many features of this theory, but dispenses with the precondition that natural kinds ought to exist in an equilibrium or possess a single, central mechanism. As such, Khalidi (2013, 2015) sets a congenially low-bar for what can count as a natural kind: it need not have an essence, or manage a system in homeostasis, but merely instantiate a set of causally related and tractable properties that, in turn, project to and license epistemically valuable inductive inferences about the phenomena we are interested in.

More specifically, in Khalidi's account natural kinds are associated with a network of causal and determinate properties that reliably generate a "rich set of effects," and that ground the generalizations that kind terms feature in, allowing us to predict and explain the relevant phenomena at hand (Khalidi 2015, 7). There is an important distinction to be made here between determinate and determin*able* properties. A determinable property is one that can range over a set of values, so for instance consider that mass, or viscosity, can encompass a wide if not infinite range of instances: an object could be 2 kg or 2×10^{30} kg, motor oil has a viscosity of around 600 cPs, and so on. Under Khalidi's account, natural kinds should pick out determin*ate* properties, that is, ones which have an expressed value or range of values associated with them. For example, "Gold" picks out an atom or collection thereof with an atomic weight of about 197, depending on the ratio of isotopes present. Properties with fixed values play important roles in generating the kinds of inductive inferences that license the natural kind term's projectibility, for instance that solid lumps of Gold, being of atomic weight ~ 197, melt at approximately 1064 °C. In contrast, failing to fix a property leaves it explanatorily empty, and so returning to our earlier case, one can't really say much about the set of entities with mass *simplicter* beyond the claims that they have mass and cannot travel at the speed of light. It's only when we fix a property that we can begin to make predictive use of it.

In what follows I argue a fundamental property common across models of working memory, its *maintenance* of information, is not determin*ate* and does not help explain or predict the central cognitive processes that we are interested in, from consciousness to reasoning. As matter of fact, the maintenance of information occurs throughout the brain in the service of most behavior. In turn, restricting working memory to a more

[4] Although whether a robust homeostasis, such as an equilibrium, is a necessary feature of HPC natural kinds has been debated (cf. Craver 2009).

tractable property—as authors have recently suggested by associating working memory solely to the *manipulation* of information—may guarantee a predictive connection with central cognition. However, this connection is achieved by *stipulation* as the tasks measuring working memory manipulation invoke *the very same* central cognitive processes, including flexible online inference-making and reasoning, that working memory is supposed to explain. This generates our eventual dilemma, where working memory either merely redescribes cognition, or straddles an explanatorily idle space between person-level central cognitive tasks and neural happenings.

4 What counts as working memory?

At the start of this paper I characterized working memory as the capacity that enables us to hold some information in the forefront of our minds for a short duration in the service of goal-directed behavior, and I illustrated its action using the well-trodden example of keeping a phone number in mind long enough to write it down. However, this description does not give us a theoretical, anatomical, or functional decomposition of just how such a capacity is supposed to work, and as such we cannot determine whether it meets the criteria of kindhood. At the same time, the psychological literature on working memory, which now spans over sixty years, does not offer a simple path forward. Cowan (2017) surveys *nine* distinct definitions in the literature, which run the gambit from Newell and Simon's first use of the term when describing their 1956 Logic Theory Machine to more recent and neuroscientifically anchored models (e.g., Engle 2002). Adams et al. (2018) additionally contrast twelve models on three orthogonal axes of comparison.[5] Exhaustively unpacking and detailing a taxonomy of views on working memory is manuscript-length project, and even a journalistic review of past trends in the literature is not feasible here (for instance, consult Miyake and Shah's 1999 volume). This forces us into the difficult position of having to choose or generate an exemplar that we can use for the rest of our analysis, ideally one that can avoid a straw-man objection.[6]

One might start with what has arguably become the "standard," or *Multicomponent,* model of working memory pioneered by Baddeley and Hitch (1974), discussed below in section 4.2, and which cameos in the opening paragraphs of most empirical papers on the subject. However, popularity isn't a guarantee of truth, and although nods to this model are still frequent, most neuroscientific research on working memory has turned away from this decidedly psychological account. Because a core claim of this paper relies on identifying the explanatory flaws ossified within working memory and particularly how they imperil philosophers' use of the term to gain leverage on central cognitive processes, it may be best to start with the accounts that philosophers themselves gravitate towards.[7] In what follows, I line up and contrast three working

[5] Both surveys naturally tend to favor Cowan's own "generic" or "long-term working memory" view.

[6] Here, I am following in the rhetorical footsteps of Michaelian (2011) who arguably had a more difficult task in charitably characterizing most other types of memory using the template of the multiple memory systems hypothesis.

[7] At the urging of a reviewer, though at the risk of preempting the conclusions reached in section 4.4, I'd like to make clear that the explanatory flaws to be identified are the functional decomposition of working memory into the maintenance and manipulation of information, alongside its wide purview over most cognitive activity.

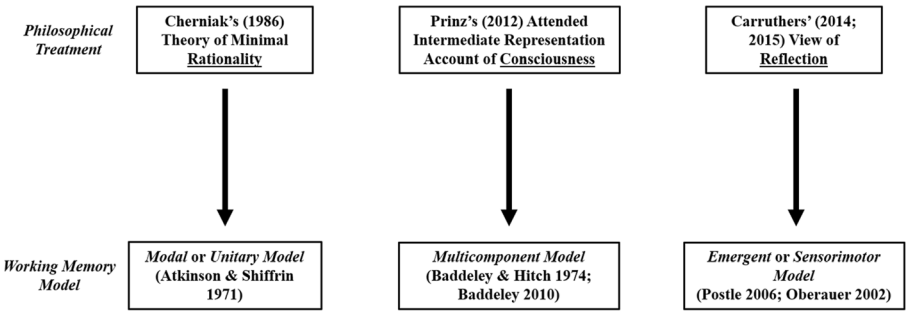

Fig. 1 The top row identifies three major theories of central cognitive processes (underlined), ranging from rationality, to consciousness, and reflection. These theories appeal to different models of working memory, including the Unitary, Multicomponent, and Emergent models, below

memory models that play a crucial role in three prominent philosophical treatments of central cognitive processes, as sketched below in Fig. 1.

A few words before we move onto the technical intricacies of these models. First, the figure above only captures a rough sketch of the dependencies between philosophical treatments and working memory models. The reality is far less clear cut, but in a way that only sharpens the polemical edge of this project. For instance, although Cherniak largely favors the language used by Atkinson and Shiffrin, he also gestures at the fragmentation featured in Baddeley's model. Likewise, both Prinz and Carruthers help themselves to aspects of many models, not all of which are compatible with one another, with Prinz often including aspects of sensorimotor accounts when moving between psychological- and neural-levels of explanation (Gomez-Lavin 2017). However, and serving as a second point, tracking the messy mapping between treatments and models isn't necessary, as all these models are doomed or saved by their common characterization of working memory's *functional role*. Finally, this is not an exhaustive treatment of these models, for a more detailed picture consult Cowan (2008, 2017) alongside Miyake and Shah (1999). Instead, the goal here is to showcase the general theoretical silhouette of working memory and to isolate a common thread that can underwrite a charitable, *generic* account that we can evaluate as a natural kind candidate.

4.1 The Unitary Model

Cherniak's (1986) account of rationality dispenses with the large-scale, maximally rational agent models that were and are still popular in formal epistemology and other agent-based sciences—ones in which, to caricature things a bit, an agent must survey all of their background beliefs before accepting a proposition in order to avoid contradiction. Instead, Cherniak attempts to deduce the limits of what could count as minimally rational, often turning to "realistic" models of cognitive processes, like working memory, to motivate his argument (52–4). For Cherniak, memory can be split between a passive, large long-term store and a "working memory" that is identified as an active, short-term store limited to six items that can be maintained, and "upon which operations can be performed, such as making deductive inferences" and whose contents "correspond to what [one] is now thinking about" (52–3). In this way, working memory

forms the arena for on-line thinking and reasoning and provides a biologically plausible limit to the kinds of reasoning processes that an agent should possess.

Though Cherniak makes no explicit mention of Atkinson and Shiffrin's (1971) Unitary Model, or any primary source psychological text, it is the only model which demonstrates all of the features that Cherniak leans on.[8] As mentioned previously, Atkinson and Shiffrin were not the first researchers to use the term working memory, both Newell and Simon (1956) and Miller et al. (1960) used the term in different ways; however, Atkinson and Shiffrin gave working memory its first *functional* characterization reproduced in Fig. 2 below.

In this model of human information processing, information from the environment is first processed by sensory systems, after which it can move to the short-term store—deemed a "working memory"—in which an array of control processes can maintain, via rehearsal, or manipulate its contents, eventually either shunting them off to a large, passive long-term memory store, or precipitating a behavioral response (Atkinson and Shiffrin 1971, 2–5). It should be apparent just how much of cognition is attributed to working memory—in effect, it serves as a bridge between perception and action. As Atkinson and Shiffrin themselves note:

> Because *consciousness* is equated with [the short-term store], and because control processes are centered in and act through the [short-term store], this store is considered a 'working memory': a store in which *decisions are made, problems are solved, and information flow is directed* (Atkinson and Shiffrin 1971, 5 emphases mine).

Notice how working memory is responsible for consciousness, decision making, problem solving, and general information flow, and this litany of cognitive responsibilities is both the reason that working memory is so attractive for researchers interested in central cognitive processes—like many contemporary philosophers of mind—and what will ultimately preclude us from finding a coherent realizer. This error could be identified as early as 1976; for instance, consider Craik and Levy's characterization of the model where they point out that "short-term store [of working memory] is much more than a simple stage of learning; it is the control center for *all cognitive activity*" (Craik and Levy 1976, 165 emphasis mine). As we will see, a commitment to working memory's wide scope is perpetuated throughout each of its iterations.

4.2 The Multicomponent Model

Prinz's (2012) Attended Intermediate Representations theory of consciousness incorporates working memory as a necessary component that enables conscious access to mental representations (92). Greatly simplifying the view, an intermediate-level representation is rendered conscious when it is boosted by certain attentional processes—particularly by top-down driven, gamma-band oscillatory dynamics—in such a way that it is made available to working memory (143, 321). While Prinz relies on both psychological and neuroscientific accounts of working memory, his theory leans heavily on two aspects of Baddeley's Multicomponent Model: executive control and

[8] Rather, Cherniak is largely pulling from textbook and reference sources, including Klatzky (1975).

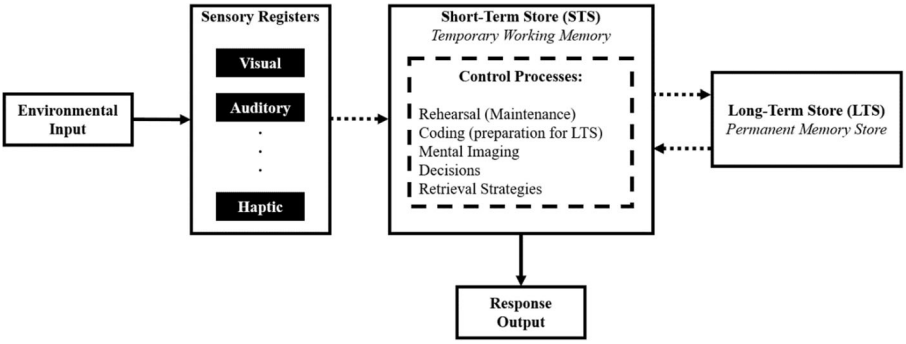

Fig. 2 A diagram, adapted from the original, of the functional relationships posited by Atkinson and Shiffrin's 1971 model of "human information processing." Bolded arrows indicate mandatory operations, while dotted arrows indicate contingent relations

fragmentation. As Prinz puts it, "working memory is a short-term storage capacity that allows for 'executive control'," and "working memory is not a single subsystem but a family of functionally analogous subsystems, corresponding to different modalities," citing Baddeley's 2007 book as a source for each of these aspects (92, 249). This emphasis is likely due to Prinz's theory being a non-central account of executive cognitive processes; particularly, top-down attention and consciousness.

The Multicomponent Model, reproduced below in Fig. 3, carries over many functional features of the earlier Unitary Model, but influenced by the results of Baddeley and Hitch's (1974) dual-task paradigm and Shallice and Warrington's (1970) survey of

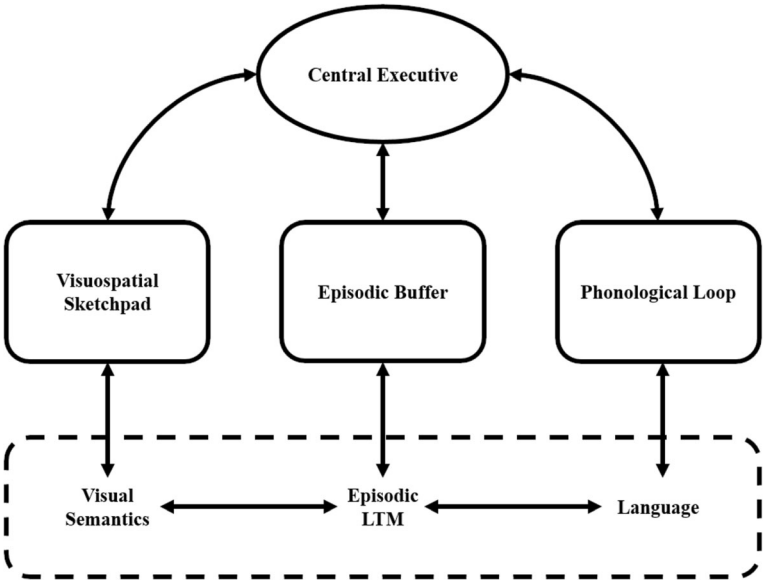

Fig. 3 A version of Baddeley's multicomponent model of working memory, a capacity comprised of the Central Executive, Visuospatial Sketchpad, Phonological Loop, and the newly added Episodic Buffer. Stored knowledge that is accessed by the relevant subcomponents of working memory is represented in the lower, dotted box (adapted from Baddeley 2010)

brain lesions, it fragments the single short-term store into two modality-specific subsystems driven by a central executive. The phonological loop enables us to maintain or rehearse phonetic elements, perhaps by inner speech, while the visuospatial sketchpad enables the maintenance and manipulation of visual information, as happens for instance when mentally rotating a 3D object from a 2D depiction on a piece of paper (Baddeley 2010, R136; Shepard and Metzler 1971). The episodic buffer was later introduced by Baddeley as an attempt to characterize how working memory could maintain and make use of multi-modal cognitions (Baddeley 2010). This generative and imaginative capacity, which Baddeley associated with consciousness, could not easily be explained by the original trifecta of the two sensory systems plus the central executive. In fact, in earlier iterations of the model there was no clear way to explain how working memory could even access the *episodic* contents of long-term or autobiographical memory, and so, the episodic buffer was posited. Finally, the central executive is tasked with the difficult work of running the show of working memory, deciding when to enlist the specialized subsystems, when to employ the episodic buffer to synthesize new thoughts, when to call on motoric or evaluatory systems, and so on. Its agency and position atop the hierarchy begets a worry of homuncularity, since it seems to be deciding what to do, and by extension what you do. Baddeley was aware of these concerns, deeming the central executive a "conceptual ragbag" and a placeholder for the complex and little understood goings-on of cognition. A resultant aim of cognitive psychology going forward should be to "sack" the central executive (Baddeley 1996, 6–9).

Though this account looks quite different from the earlier Unitary Model, it carries over the same wide scope functional commitment as before. Consider how Baddeley notes that these systems, when working in tandem, "are assumed to be necessary in order to keep things in mind while performing complex tasks such as reasoning, comprehension and learning," and working memory does so as it, "provid[es] the ability to maintain and manipulate information in the process of guiding and executive complex cognitive tasks," (Baddeley 2010, R136; Repovš and Baddeley 2006, 5). Working memory, through the processes of maintenance and manipulation, still realizes an arena for central cognition.

4.3 Emergent Accounts

Carruthers (2014, 2015) homes in on working memory as the empirically vetted realizer grounding philosophers' longstanding commitments to the mind's "central workspace" that allows for kind of flexible reasoning and reflection that features prominently across central cognitive processes (Carruthers 2014, 144). As he puts it, "there is, indeed, a central workspace in the mind whose contents are always conscious. This is so called 'working memory'" (145). Though Carruthers helps himself to aspects of a variety of models, for his account of reflection and consciousness he largely settles on a neurally-anchored view of "working memory as a sort of virtual system that uses executive resources located in frontal and parietal lobes to direct attention towards mid-level sensory areas, issuing in images that can be sustained, transformed, and manipulated in ways that are globally accessible," (Carruthers 2015, 146). Though it's not quite clear what a "virtual" system means in this context, the notion that working memory's primary functions of maintenance and manipulation are realized by interactions between frontal "executive" brain regions and representations located in sensory

areas is consistent with Emergent or Sensorimotor accounts of working memory that have gained significant traction in cognitive neuroscience over the last decade.

Although Emergent accounts resemble earlier minimalistic psychological working memory models—particularly Cowan's long-term, or "generic" working memory view (Cowan 1988)—they differ by being primarily driven by neuroscientific considerations. Postle's 2006 manifesto, challenging what were then the "standard" prefrontal accounts of working memory, lays out a radical revision, arguing that working memory is better understood as a massively fragmented process that emerges from the interaction of frontal and sensory cortices. Prior to this, the standard view in neuroscience—motivated by the multi-level neuroscience of Goldman-Rakic (1995) and the electrophysiological work performed by Funahashi et al. (1989)—held that activity in the prefrontal cortex was necessary for the maintenance and manipulation of working memory contents.[9] Postle marshals a diverse range of evidence suggesting that prefrontal cortex is not strictly necessary for working memory performance, and instead argues that wherever representations can be created in the brain, there they can be maintained and possibly manipulated (10). The tagline is "if the brain can represent it, the brain can also demonstrate working memory for it" (10). How does this come about? According to Postle, "working memory functions are produced when attention is directed to systems that have evolved to accomplish sensory-, representation-, or action-related functions…working memory may simply be a property that emerges from a nervous system that is capable of representing many different kinds of information" (10). Carruthers adopts this language to buttress his own account of reflection and consciousness.

In the fourteen years since its introduction, many brain regions have been found that maintain stimulus specific information during working memory tasks. Christophel et al. (2017) produced an excellent summary of where this stimulus specific activity can be found across the Human and Macaque brain, which is reproduced in Fig. 4 below.

Here we can see that Postle's tagline bears some truth: Where representations are created, there they can be maintained and manipulated. And this realization makes intuitive and evolutionary sense, why shunt representations over to a specialized brain area for safe keeping when that only doubles the neural work required? One might even count this move away from functional boxologies and towards a neurally-constrained and distributed model of working memory as genuine progress. However, peeling apart the exciting contributions of neuroimaging to cognitive neuroscience, we find an eerily similar functional picture at the heart of even this radical revision to working memory. At the conclusion of their survey, Christophel et al. (2017) echo the same expansive, wide scope functional role with which working memory was identified nearly fifty years ago:

> There is abundant evidence for widely distributed stimulus-related information in sensory, parietal, and prefrontal cortices during working memory delays… we suggest that working memory is better characterized as a distributed network *that*

[9] Specifically, during the *delay period* common to working memory task paradigms. In most working memory tasks, a target is presented to be maintained or manipulated, followed by a delay period in which the target disappears, and during which time distractors may be introduced. Afterwards, an additional target or response probe is introduced.

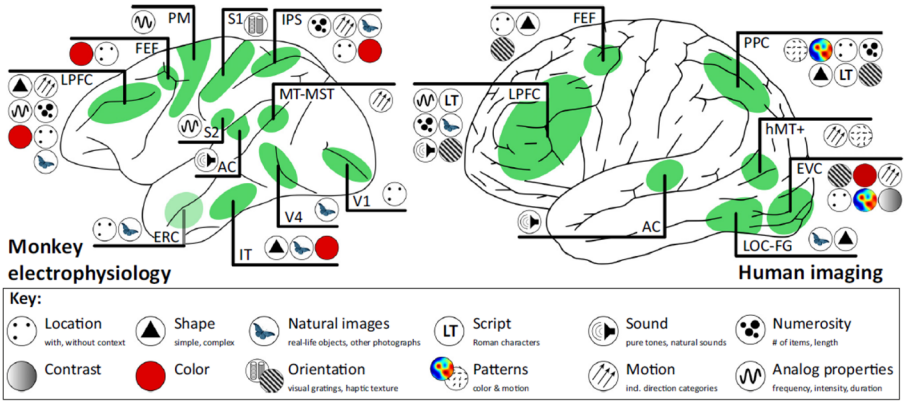

Fig. 4 From Christophel et al. 2017, this figure collects a series of neuroimaging and electrophysiological results from the Macaque (left) and Human (right) cortex. We see that a variety of stimuli, including abstract features associated with graphemes and numerosity, are represented throughout the cortex. These findings have led the authors to propose that working memory is distributed throughout the brain

> *gradually transforms sensory information towards an appropriate behavioral response*, across a temporal delay…This notion suggests that perhaps the field of working memory should shift its focus from asking where in the brain working memories are stored to *unraveling how a range of highly specialized brain areas together transform a sensory stimulus into an appropriate response* (120, emphases mine).

After all, identifying working memory as a system that processes and transforms sensory information into a behavioral response resembles the model of human information processing offered by Atkinson and Shiffrin (1971), merely substituting a functional box for a functional cortical mapping. Stepping back a bit, understanding how we move from perception to action is largely *the same project* as understanding cognition, and this realization foreshadows our ultimate conclusion that working memory merely redescribes the thing it's supposed to explain, that is, cognition.

4.4 Towards a generic model of working memory

Before we move on, there are three important things to note about our detour through these models of working memory used by philosophers. First, there is a distinct trend towards *fragmentation*. Where we began with a Unitary store, which Baddeley broke apart, now it appears that the brain may have dozens or even hundreds of working memory stores and subcomponents. Of course, part of the motivation for reframing working memory as an *emergent* property is to avoid generating a model with scores of components or the bloated boxology that would be required to capture it. Looking past this metaphysical dodge for a moment, what began as a single system now looks to be something realized across the brain, and this fact alone—besides the theoretical and anatomical disagreement amongst models of working memory—should drive intuitions against the view that it is a good natural kind candidate. However, and this leads to the second point, despite the different levels of description in the models we've surveyed, the functional core of working memory as an expansive system reliant on maintenance

and manipulation of information and implicated across cognition, continues throughout iterations in its fifty-year history. Finally, and this will play a role in the following sections, because of its seeming ubiquity throughout the cortex and across cognitive tasks, some scholars of the emergentist-bent—including Postle (2016)—have reintroduced a distinction between short-term and working memory, restricting the latter to *manipulation* or the "flexible use" of mnemonic contents (44). Why this is important will become clear shortly.

Is there a common thread that ties these models together? I hope it is now evident that they share a functional heritage that has shadowed working memory research since its inception: working memory is realized by the *maintenance* or *manipulation* of some information, no longer in the environment, for limited durations in the service of goal-directed behavior. Call this our *generic account* of working memory. Everything we've surveyed is consistent with and depends upon this functional characterization, which can also be found throughout the definitions reviewed by Cowan (2017) and Adams et al. (2018). More importantly, each of the philosophical accounts of central cognitive processes that depend on working memory lean on it precisely for its functional abilities to hold contents in mind and to transform them.

5 Why working memory is not a natural kind

Some might already have begun to feel pessimistic about working memory's candidacy as a natural kind term; after all, with little in the way of theoretical or anatomical agreement, and tied to functions that may be multiply realized across the brain, it's already doubtful that we could pick out a coherent set of causal properties that reliably project to the phenomena of central cognition. This worry is sharpened by the twin functions of *maintenance* and *manipulation* at the core of our generic account. Recall that Khalidi's congenial approach to natural kinds requires a set of determin*ate* properties, ones that are tied to a value, whereas *maintenance* and *manipulation* stand remarkably underspecified. Most any neural structure could be said to maintain or even manipulate information, from an individual neuron's membrane potential to the potentiated synaptic connections that underlie long-term memory, and so it appears that our generic account cannot even meet the minimal criteria to be considered a kind candidate.

However, this would be an unfair characterization of the working memory literature, as it's not just *any* maintenance that counts, but in fact many efforts have been made to render this property tractable and, in an important sense, determin*ate*. Going through every one of these constraints is not feasible here, as there are almost as many operationalizations of these terms are there are papers on the subject—operationalizations that for instance, specify the number of items to be maintained, or the time limit that they should be maintained, and so on. Rather, what is required is an argument to show that any causally tractable characterization of these functions fails to pick out a coherent set of features whose operations reliably project to central cognitive phenomena, and thus whose operation fail to license the kinds of inductive generalizations used to predict and explain central cognition. Setting aside stipulative or vague constraints on these functions—e.g., the maintenance of information for three or four or five seconds, etc.—which invite a Sorites paradox, we can focus on a proposed causal

realizer of working memory maintenance that has driven most neuroscientific research on the subject and which we've already introduced: *delay period activity*. Let's call this delay period activity associated with working memory task performance "*maintenance-c*" to indicate that it is the causal candidate realizer of working memory maintenance. The argumentative steps that we will proceed through are as follows:

(1) Many brain regions demonstrate *maintenance-c*
(2) Neural firing associated with *maintenance-c* occurs outside working memory tasks, and may be supported by several mechanisms
    (C1) Hence, *Maintenance-c* does not isolate a univocal neural signature associated with working memory
(3) Working memory requires an even stricter constraint, *manipulation*
(4) *Manipulation* either reduces to *maintenance-c* or it circularly appeals to central cognitive processes
    (C2) Hence, either working memory encompasses the many ways information is held and transformed in which case it cannot be a natural kind, or working memory, by invoking the very processes it is meant to explain, sits explanatorily idle

Before we go through the argumentative steps in order and conclude, I want to preempt two possible objections to the strategy laid out above: namely, what might be criticized as an unreflective appeal to neuroscience in our search for the realizer of working memory, and, the ongoing slippage between the terms of *maintenance* and *manipulation*.

First, we're turning to neuroscience here as it offers the most readily *causal* candidate realizer of working memory via delay period activity. As mentioned earlier, psychological and computational models are often less concerned with the messy details of implementation. As such, there may be space for a purely functional characterization of working memory, perhaps in phenomenological or person-level terms and similar in kind to Fernandez' functional account of episodic memory (Fernandez 2019).[10] Though such a person-level depiction of working memory may be possible, it would also be highly unorthodox, as in all the models reviewed working memory is treated as a sub-personal system whose operation *occasions* person-level mental life, including consciousness and deliberation. Furthermore, it is precisely this special connection forged between sub-personal happenings and rich person-level cognitions that anchors philosophers' appeal to working memory in their theories of mind.

To the extent that any functional characterization relies on implicit, or makes explicit, claims about the *sub-personal* boundaries or operations of working memory, it will be subject to similar scrutiny as our generic account introduced above. For instance, if working memory is identified with a sub-personal system that carries out the function of maintaining information, then we can presumably ask a number of questions—including, just what kind of information is maintained, for how long, by which processes, etc.—whose answers shape a story about its implementation. Evaluating this implementation, then, requires that we render the functional properties under

[10] I'd like to thank an anonymous reviewer for pressing me to expand on this point.

investigation, such as maintenance, determinate by specifying or operationalizing them, as we've done with *maintenance-c*. By rendering properties determinate and tying their implementation to a causal realizer, we can then use the framework of natural kinds to evaluate their explanatory leverage over the phenomena that interest us.

Second, I want to acknowledge a slippage between the terms of *maintenance* and *manipulation* that has been present in this paper so far. This slippage largely reflects the lack of a coherent or consistent conceptual division amongst these functions in the literature, at least until recent attempts by LaRocque et al. (2014) and Postle (2016), among others, to restrict working memory to an account of manipulation. For most of working memory's history maintenance and manipulation were brought up together, and this makes some sense, as prima facie they just unpack the term itself. When we consider that "working memory" is a dyadic construction that contains the concepts of holding something in mind in order to do something to or with it, such a slippage seems less pernicious. I have chosen to begin with maintenance not only because every model and every task surveyed requires keeping some information in mind, but because beginning here will help us better expose the circularity that appeals to *manipulation* smuggle in.

5.1 Many brain regions demonstrate *maintenance-c*

Maintenance-c singles out a causal candidate realizer of working memory; that is, the increased neural activity found during working memory tasks that has been hypothesized to encode stimuli during the task's delay period where they are no longer in the subject's environment, or in other words *delay period activity*. Its roots go back to early electrophysiological work conducted on monkeys by Fuster and Alexander (1971) and Kubota and Niki (1971) who both found increased activity in prefrontal neurons during the delay period of a simple working memory task.[11] In a follow-up electrophysiological study, Funahashi et al. (1989) found that the delay period activity of select prefrontal neural-populations encoded the spatial location of a previously presented stimulus (334, consult their Fig. 3 for an example). Much of this work was codified by Patricia Goldman-Rakic (1995), helping to usher in the *prefrontal dogma* of working memory, which held that working memory is realized by the activity of prefrontal neurons whose role it is to encode the contents to be remembered.

This prefrontal dogma is the same one challenged by Postle et al. (2006) and the rise of Emergent accounts reviewed in section 4. Postle et al. (2006) cites several studies in both monkeys and humans showing proficient working memory performance despite significant lesions to areas of prefrontal cortex. For instance, Petrides (2000) lesioned the dorsolateral prefrontal cortex of his primate subjects and found that they performed similarly to controls (7500). Additionally, D'Esposito and Postle (1999) reviewed data from 24 patients with severe prefrontal lesions and found minimal deficits to working memory task performance amongst this set.[12] These results demonstrate that prefrontal

[11] Here, the monkeys were shown a piece of fruit that was placed in one of three locations, they had to retain that information while a screen was lowered, and only after a delay were they tasked with indicating the location of the fruit (Kubota and Niki 1971, 338).

[12] As a caveat, these are very rudimentary tasks, often only requiring the maintenance of three or four stimuli, and other deficits, particularly in monitoring the task, were correlated with prefrontal damage; however, the tasks are identical to those that established the prefrontal dogma of *maintenance-c*.

cortex is not strictly necessary for working memory performance, and so it invites us to ask where *else* can we find the delay period activity we've associated with *maintenance-c*.

As was discussed in section 4, recent neuroimaging data support the Emergentist hypothesis that wherever representations can be crafted in the brain, there they can be maintained during a delay period (Christophel et al. 2017, consult Fig. 4 in section 4.3). Again, this makes intuitive and evolutionary sense: there is no need to re-encode the contents to be remembered in prefrontal cortex when they have already been encoded elsewhere. The takeaway is simple, it appears that delay period activity, and hence *maintenance-c* is a generic property common throughout the brain.

5.2 Neural firing associated with *maintenance-c* occurs outside working memory tasks, and may be supported by several mechanisms

Despite the presence of delay period activity, and hence *maintenance-c,* across the cortex, perhaps every bona fide case of *maintenance-c* that realizes working memory task performance is united by a univocal neural mechanism. This mechanism could then ground working memory's natural kind status. Which neural mechanism is associated with *maintenance-c*? As we've already covered, *maintenance-c* is tied to delay period activity, and this activity is just increased rates of neural firing as was first discovered by the electrophysiological studies conducted on monkeys reviewed above. Now standard neuroimaging tools, including fMRI and EEG, which respectively measure metabolic—via blood oxygenation levels—and electrical—via population level neural oscillatory dynamics—changes correlated with increased rates of neural firing, are also used as less invasive proxies to detect large scale changes in neural activity (Wang et al. 2016; Gevins et al. 1997).

The ubiquity of neural firing dynamics throughout the nervous system begins to outline the contours of a dilemma for any proposal that identifies these dynamics as the realizer of *maintenance-c*. Under a broad interpretation where we identify *maintenance-c* with increased neural firing simpliciter, we are guaranteed to find instances of *maintenance-c* throughout the brain—and indeed the entire nervous system—which are not associated with working memory task performance and thus which do not project to, explain, or predict the central cognitive phenomena that working memory is supposed to underwrite. Consider that increased neural firing is associated with perception, learning, long-term potentiation and memory formation, motoric control, and just about every single cognitive process one can think of, and features in pathological cases including, epilepsy and other disorders (Gerstner et al. 1997; Polonsky et al. 2000; Dragoi et al. 2003; Jiruska et al. 2010). Perhaps, in some cases this firing represents the maintenance of information, but it does so in such a shallow sense that it cannot license any interesting generalizations about the robust cognitive phenomena that concerns philosophers and us here.

On the other hand, stipulating that *maintenance-c* only corresponds to neural firing occasioned by working memory task performance cannot guarantee that *maintenance-c* picks out a coherent causal set of features either. Such a stipulative move depends on the assumption that working memory tasks themselves form a coherent and unique set. Another way to put it, imagine that you only had a visual test to determine whether a

metal is gold that depended upon it having a shiny and brassy colored luster. Both pyrite and gold share this feature and if one assumed that the visual test provided enough evidence to group entities into a kind, one would mistakenly group pyrite and gold together.[13]

However, the picture is still yet worse for *maintenance-c*. A meta-analysis of 189 fMRI studies on working memory, attention, and motor intention tasks show similar patterns of increased neural activity recruited by these three tasks, which intuitively and from a third-person perspective seem like distinct capacities (Rottschy et al. 2012; Ikkai and Curtis 2011 show similar results). This result was replicated by a team who trained a linear classifier on brain wide patterns of neural activity associated with these three tasks—via multi-voxel pattern analysis—and found that the classifier, when trained on one task, e.g., attention tasks, could successfully decode representations from the other two (Jerde et al. 2012). These results suggest that whatever behaviors we have intuitively group together as satisfying working memory tasks may in fact place similar demands on the brain as other capacities, including covert attention and motor intention, which we assume to be distinct from working memory.

Returning to working memory tasks, recent neuroimaging results from Lewis-Peacock and colleagues (Lewis-Peacock et al. 2012; Lewis-Peacock et al. 2015; LaRocque et al. 2014, and LaRocque et al. 2015) have raised the possibility of the *activity silent* maintenance of information. While these results are both fascinating and troubling for neuroimaging, an extensive review of the paradigms employed, and their replications, lies outside the bounds of our present project. As a cursory summary, the data suggest that the current state of our neuroimaging technologies may not be sensitive enough to detect the many ways that neural populations encode information, even in standard working memory tasks. Several replications and extensions of this work have been conducted that only cement the worry that our current tools may miss out on little understood but important neural dynamics. As Stokes (2015) summarizes, "accumulating evidence suggests that persistent delay activity does not always accompany [working memory] maintenance" (394). This pessimism is echoed by Sprauge et al. (2016) who conclude that, "[t]hese results challenge pure spike based models of [working memory] and suggest that remembered items are additionally encoded within latent or hidden neural codes that can help reinvigorate active [working memory] representations" (694**)**. What these "latent" or "hidden" codes might be, we don't really know. They could be sub-firing-threshold dynamics, short-term changes to synapses, or even calcium kinetics at the synaptic junction (consult Stokes [Stokes 2015, 396] and their Box 1 [402]).

5.2.1 *Maintenance-c* does not isolate a univocal neural signature associated with working memory

The evidence provided in sections 5.1 and 5.2 suggests that there is no univocal neural signature that corresponds to the maintenance of information in working memory tasks, which we've termed *maintenance-c*, regardless of whether it is constrained by delay

[13] The jadeite/nephrite example is a classic case of the same phenomenon in action (Hacking 2007).

period activity or increased neural firing. Delay period activity is a generic feature displayed across brain regions, and the increased neural firing that was long associated with this activity occurs throughout the nervous system. Worse yet, it appears as though the patterns of firing associated with working memory tasks are similar to a range of other cognitive capacities, and that many mechanisms, including ones that we have yet to understand, have a hand in the maintenance of information. Stepping back a bit, it is not surprising that much of the nervous system can maintain information over time, and that such a complex system possesses many mechanisms that instantiate this process. Maintenance is merely a currency of cognition, and any information consuming system that exists over time will likely share a similar functional profile.

5.3 Working memory requires an even stricter constraint, *manipulation*

Recent neuroscientific work has sharpened the distinction between working memory's twin functions of *maintenance* and *manipulation*. Again, for most of its history these functions were brought up simultaneously as they help to operationalize the dyadic nature of working-memory. At the start of their paper on short-term memory, LaRocque et al. (2014) stipulate that, "Short-term memory (STM) refers to the capacity-limited *retention* of information over a brief period of time, and working memory (WM) refers to the *manipulation* and *use* of that information to guide behavior" (1, emphases mine, parenthetical initialisms theirs). Contrast this to an earlier paper by D'Esposito and Postle (1999) that attempts to tease apart these functions by "comparing a working memory condition that required retention of information (*maintenance*) during a delay with a condition that also required the transposition (*manipulation*) of information being held in working memory" (68, emphases theirs). LaRocque and colleagues are marking a conceptual shift, one that is defended by Postle (2016): "[working memory] differs from [short term memory] in that the former entails operations that *transform* the remembered information, or that require *control processes* beyond its simple [short-term retention]. A real-world example of the former is performing mental arithmetic on the remembered total on a restaurant bill" (44, emphases mine). Continuing, Postle states that working memory tasks, "require additional cognitive operations that entail the manipulation of the information that is being retained, and/or the flexible use of that information to guide behavior in tasks that are more complicated than simple recognition or recall" (44). To summarize, for Postle and colleagues bona fide *manipulation*, or working memory, requires the "flexible use" of information for behavior, the "transformation" of information, or the operation of "control processes."

Additionally, Postle (2016) identifies three broad categories of bona fide working memory tasks that showcase manipulation: (1) Delay tasks that "require some mental transformation" of remembered contents, for instance taking a random string of letters and alphabetizing it, (2) Continuous tasks where the remembered contents must be updated; and, (3) dual-tasks, such as those used by Baddeley and Hitch (1974), which may require switching between two sets of remembered items (44). Is this quick move away from *maintenance* enough to secure working memory's candidacy as a kind term? Perhaps, but as I show below, it does so only in a stipulative and profoundly unexplanatory way.

5.4 *Manipulation* either reduces to *maintenance-c* or it circularly appeals to central cognitive processes

Let's begin with what could count as the "flexible use" of information for behavior. Take a simple maintenance task, like the one assigned by Kubota and Niki (1971) in which monkeys had to keep the spatial location of a piece of fruit in mind, and demonstrate that they knew the location by pressing a corresponding button, in order to receive their reward.[14] On the surface, this task would be excluded by Postle (2016) as a genuine working memory task; however, consider that the monkeys aren't completing the task for fun, they're keeping the information in mind to earn their reward. It's as though the monkeys were using the information to guide their behavior. Perhaps this does not count as *flexible* use; however, then the burden is on Postle to generate criteria for this kind of use, ideally ones that are not post hoc or obviously question begging. The notion of flexible use echoes how early working memory models explicitly stated that the capacity was under "voluntary control" (Atkinson and Shiffrin 1971, 2). At the same time, voluntariness, or *deciding* what to do when, is precisely the kind of central cognitive processes that is supposed to emanate from the proper function of working memory, as happens for instance when one is weighing their options. Dual-task paradigms that require switching between two sets of memoranda merely smuggle in the same central cognitive process of *decision making* while doubling the maintenance required.[15]

Moving on to the "transformation of information," again it is not clear why this could not also occur under the guise of more rudimentary maintenance tasks. Stepping back, when we hold something in mind, like a phone number, the phenomenology of our internal representation shifts. It's not as though we still have the number in front of us, rather it has morphed into a representation amenable to internal speech.[16] But it is likely that Postle (2016) means something closer to the continuous task paradigm, in which, for instance, you are given an initial set of stimuli to keep in mind—say, three numbers: 4, 1, 7—and as the task progresses you are asked to "transform" the numbers by addition or subtraction, so in our example you would see a second screen with the following: +1, _, −2. First off, it's not immediately clear that working memory is *transforming* the remembered contents, rather in these cases one is calling upon inference procedures—such as the rules governing addition and subtraction—to generate a new representation to be remembered. But recall, working memory is supposed to explain how we can instantiate central cognitive processes, which include inference and decision making and general problem solving.

The problematic invocation of "control processes" as a guide to discover bona fide instances of working memory is made clearer by the delay tasks that Postle (2016) cites and which feature in his work with D'Esposito et al. (1999). In these tasks, one is given a jumbled string of letters, say XBTLD, and then in half the trials one must internally reorganize these letters along a rule set—for example placing them in alphabetical

[14] Recall that D'Esposito and Postle's (1999) earlier work debunking the prefrontal dogma of working memory featured similar paradigms in humans.

[15] And it appears that Chimpanzees are also capable of successfully negotiating dual-task conditions (consult Völter et al. 2019, especially experiment 2).

[16] Cf. Hume's distinction of the force and vivacity brought on by impressions versus ideas for a similar intuition.

order—and then one must compare that internal representation to a probe shown at a later point in the trial. So, for instance, at a later point you might be shown a number, e.g., "4," and then shown a target letter, e.g., "T," and asked whether the letter shown matches the correct letter in the reordered string (in this case, it would). I should make it clear, though the task seems trivial it is in effect demanding *a lot* of cognitive work: One has to appreciate and follow the instructions, one must understand which condition they're in and how that is supposed to affect their responses, one must attempt to retain a string of letters that are very briefly presented, reorder them according some rule set, process an ordinal quantity, compare this to an internal representation, and decide on an appropriate response, all the while—and I can say this as a former subject in similar experiments—wondering when this will be over and just which subway will be the least worst choice to take home. These tasks represent a tremendous amount of cognitive work—perhaps more than most people do throughout their regular day—that we're compressing into a test for a single capacity, when ideally all the contributing cognitions should be parsed.

Why does the introduction of "control processes" pose a challenge to working memory research? After all, consider that even in Atkinson and Shiffrin's early model, the box housing working memory contained a list of *exactly* that; namely, control processes. Here's where conceptual bookkeeping and clarity are especially important. In the Unitary model, only *one* control processes is detailed, that is, how contents stored in working memory are rehearsed, and this was "modeled" mathematically and described using the computer-like metaphor of "slots" in which items are continuously refreshed (Atkinson and Shiffrin 1971, 9). The other processes are left unmentioned, except by featuring in the definition of working memory as a system that is responsible for decision making and problem solving (5). There is room for an important distinction here between candidate *mechanisms* that support working memory and just the kinds of person-level behavior working memory is itself supposed to underwrite. Rehearsal is a potential mechanism that helps explain how we keep somethings—usually phonetic stimuli—in mind, and it reappears in Baddeley's phonological loop. However, "mental imaging" or "decision making," as Atkinson and Shiffrin put it, are not proposed mechanisms, but in fact are the kinds of person-level behavior that working memory is supposedly responsible for. As such, if by "control processes" what Postle (2016) meant were proposed *mechanisms* for how working memory does whatever manipulation is supposed to be, then we might have to give him some leeway. But instead, these control processes are just undefended operationalizations of the kinds of cognition working memory is supposed to explain.

Taking stock, *manipulation* is not a coherent functional feature implicated across working memory tasks, rather it is something measured by adjacent cognitive tasks that themselves require problem solving and decision making or by gut-appeals to a task's "flexibility" or utility. Ultimately, it is a stipulative move to restrict what counts as a working memory task, thus separating the capacity from the troves of data implicating it, via *maintenance-c,* across cognition and the brain.[17] I have shown that tasks testing *maintenance-c* meet the minimal bar of manipulation in terms of "flexible use," as maintenance does not occur for its own sake, but for planning and enacting future

[17] In fact, this is made explicit by Postle (2016) who stipulates that working memory must require prefrontal control, and so tasks that fail to recruit the prefrontal cortex should not count as proper working memory tasks (45).

behaviors. A stricter reading of manipulation requires the invocation of more complex cognitive procedures, including reading comprehension, decision making, problem solving, and reasoning, which themselves are the kinds of central cognitive processes that working memory is supposed to *explain* (Jonides and Nee 2006).

5.4.1 Hence, either working memory encompasses the many ways information is held and transformed in which case it cannot be a natural kind, or working memory, by invoking the very processes it is meant to explain, sits explanatory idle

The form of this dilemma has been long in the making: Either we accept that working memory is tied to *maintenance-c* of information, in which case it is something that is realized throughout cognitive processes and is present across the brain in a plurality of ways, or we artificially restrict working memory to a set of tasks bound to the central cognitive processes that its operation was expected to explain. Following either horn working memory will fail to pick out a coherent set of casual properties whose instantiation reliably project to the processes that working memory is supposed to sustain; that is, those central cognitive processes of rationality, deliberation, and flexible thought assumed to emanate from our mind's internal workspace that allows thoughts from across the mental economy to congregate and be synthesized with one another.

If working memory is realized by *maintenance-c*, then it is instantiated by many mechanisms—some of which we have yet to fully understand—across much of the brain the service of most robust cognition. And this makes some intuitive sense, after all, consider how most cognitive activities require keeping information in mind. Even psychological experiments that test other capacities, like metacognition, require that you learn the cues and the task parameters and that you keep these in mind throughout the task's duration. In this broad sense, working memory is just what we do cognitively most of the time—and we shouldn't be surprised by this ubiquity as it has been present in the theoretical and functional framing of working memory since its inception.

If we restrict working memory to tasks measuring "control processes" that invoke problem solving and decision making, then we may have secured a kind of projectibility between whatever realizes these behaviors in the brain and the behaviors themselves, but it is one achieved by stipulation and in which working memory plays no explanatory role. It is reasonable to try to understand how we go about solving problems and making decisions, but perhaps we should study these behaviors independently, rather than packaging them together within an omnipresent and overburdened construct like working memory that may in fact obfuscate important distinctions between these behaviors and how they are variously realized in the brain.

6 Tying up loose ends and what this means for working memory and cognitive science going forward

This project has taken us a lengthy way, through deep dives into the psychological and neuroscientific literature and into the philosophical realm of kindhood and explanation. I want to take a few moments to summarize the impact of the project, dispel some concerns, and gesture at what comes next. Working memory is a cornerstone of contemporary cognitive psychology that should explain how we keep information in

mind and transform it to suit our behavioral needs. It has begun to play an important explanatory role in philosophical theories about cognitive behaviors that have interested philosophers of mind for centuries, including reasoning, consciousness, and reflection. What I have shown is that working memory can play this role because it captures something fundamental and common to cognition; the fact that as information using and consuming systems, we need to keep things in mind. At the same time, we shouldn't be surprised that such a fundamental function is realized in many ways, such that it fails to meet even a low bar of natural kindhood. Rather than unite the diverse ways we can hold and transform information in the service of cognition behind the terminological veil of working memory, we should actively try to understand this multiplicity. Tracking just how different neural systems at various levels hold information may give us purchase on explaining the ways in which creatures like us consume and use information throughout our cognitive lives.

Some may worry that the rhetorical strategy demonstrated is too far reaching, extending to other cognitive categories, such as learning or attention, or to ostensibly non-cognitive aspects of the mind, like perception and sensation.[18] After all, if there is no mentation that meets the mold of kindhood, why worry about working memory's deficits on that score? First, we'd do well to consider when it's appropriate to use the argumentative apparatus of natural kinds. If a cognitive category trades in functions that are determinate and traces their implementation to neural structures, then we can ask if the operation of these realizers projects to and licenses inferences made about that category. This strategy is likely to exclude perception and sensation, at least at that initial and broad level of description, as it's not clear one could argue that all perception or all sensation respectively have a unique functional profile that is specifically realized.[19]

But if we were to grant that perception or sensation, or some cognitive category, is a contender for natural kindhood, whether it is subject to the same treatment as I gave working memory in this paper rests on its position in the 'explanatory stack.' Consider perception: intuitively it cleaves apart a domain of inquiry, tools, methods, and tasks from other areas in the study of the mind. But what does perception *qua capacity* explain? Which higher mental attributes, processes, or mysteries hinge on a coherent, functionally articulated, causally tractable account of perception? As one fills in that side of the explanatory divide, then one concurrently raises the explanatory pressure placed on such a view of perception and thus makes evaluating its candidacy as a natural kind more pressing.

This is what makes the study of working memory particularly germane: it straddles an explanatory position between a detailed description of neural happenings and the central cognitions that interest philosophers and psychologists. What I have shown in this paper is that the functions tied to working memory are realized across the cortex, likely through a variety of mechanisms, and are common to several cognitive operations.[20] Again, it's even

[18] I'd like to thank a reviewer for persuading me to reflect on these issues.

[19] That is, it's not clear that these are treated as natural kinds or natural kind candidates, though they might be thought of as property clusters anchored by representational (e.g., by using primarily analog or iconic representations: Quilty-Dunn 2019), computational (e.g., by featuring aspects of modularity: Firestone and Scholl 2016), or behavioral (e.g., by reverse-engineering the paradigms and tasks used to test perceptual capacities in a similar way to Buckner's 2015 treatment of cognition) features

[20] And, conversely, that artificially restricting working memory to a post hoc subset of tasks (e.g., manipulation tasks) will leave it unable to project to the broader class of central cognitions that it is thought to support.

hard to conceive of cognitive operations that wouldn't involve the maintenance and manipulation of information for some purpose; such functions lie at the heart of what it is to be a computational, information consuming system that exists over time. Thus, when scholars suggests that the operation of working memory explains how it is that we can engage in deliberative reasoning (e.g., dual system accounts of reasoning akin to those proposed by Evans and Stanovich 2013), or how we can have access to the contents of our conscious thoughts (i.e., Prinz's 2012 AIR theory of consciousness), or how our thought comes to have its domain general and flexible nature (e.g., Carruthers 2014, 2015), they are all in part correct: these processes depend on the continuous maintenance and manipulation of information. However, there is nothing holy or special about these functions, they—realized in myriad ways—are merely some of the things that brains do. If this argument is correct, then tokening working memory as part of an explanation of some central cognition amounts to little more than saying that the brain, through its implementation of these commonplace functions, is responsible. Analogizing it to the example above, perception would be similarly troubled if it were tokened as a satisfactory explanation for why we see or hear.[21]

Turning to speculation for a moment, what should we do if none of our supposedly explanatory cognitive categories—from learning to memory—pick out genuine natural kinds?[22] What if it turns out that cognition itself is not a natural kind (Allen 2017) or merits the label only in a thin or honorary fashion (Buckner 2015; Ramsey 2017)? Would Fodor's much maligned *First Law* of the nonexistence of cognitive science be vindicated (Fodor 1983)? Pragmatically and in the short-term, not much would change. Cognitive scientists would continue to build models and run experiments just as baristas continue to brew espresso without worrying too much about its metaphysics. However drab such a future appears, I also believe it harbors the possibility for a *productive pessimism* about the nature of cognition, a pessimism that philosophers of science are particularly apt to administer. That is, we may be able to both identify the flaws within and resuscitate certain cognitive terms by zeroing in on and evaluating the explanatory *role* and *directionality* that cognitive categories play.

Consider again how this potential scenario for cognitive science unfolds: by the failure of our cognitive categories to *explain*. What would they—by failing to meet the epistemic bar set by natural kindhood—subsequently fail to explain? In the case of working memory, it fails to explain how the domain-general flexibility on which central cognitions depend arises from the maintenance and manipulation of information. Instead, when you token working memory as an explanation of these central cognitions you merely restate the presumed dependency between the cognitive stigma of flexibility and the twin functions of working memory. So, what is *productive* about this exercise? If anything it seems deflationary or eliminativist, perhaps encouraging us to dismiss fifty years of research and experimentation. However, such a move would be presumptuous. While working memory doesn't explain how those cognitions higher up the 'stack' come about, it does group something intuitively fundamental to cognitive processes: that they require the maintenance and manipulation of information. A key insight of the research laid out above is that there are *many ways* brains

[21] At the same time, the explanatory landscape pictured above should help identify other cognitive categories, particularly attention, intention, and executive function, which share working memory's position in the 'explanatory stack' and may be subject to a similar argumentative strategy.

[22] Again, thanks to a reviewer for raising this concern.

carry out these functions. What we might do, then, is to scavenge the literature to fill in a mosaic-like picture that captures the multiplicity of maintenance and manipulation as they're realized throughout the brain.[23] The resulting picture would not be one of unity and conceptual cohesion, and would certainly be of little use to anyone appealing working memory as a one stop shop for cognition—especially in philosophical theorizing—but I expect it will offer a more accurate and useful guide in revealing how creatures, including us, think.

At the same time, resuscitating the productivity of these cognitive categories requires a pessimism about their explanatory force, especially when these categories aim to shed light on features underlying flexible thought. Recall that scholars are attracted to working memory because it seems to provide a scientifically vetted realizer for central cognition, and in doing so it helps explain the flexibility integral to thought (Carruthers 2014). What I have shown is that there is no system—whether "virtual," real, or distributed—that realizes these functions, instead it is just what brains do much of the time in the service of behavior, and that we should relish and study this multiplicity. Though, at the limit of inquiry, we may yet discover a purpose-built realizer of central cognition, it is more likely that to make progress we will have to set aside the search for the realizer of flexible thought. This may require that we radically reframe the study of central cognition as one that understands its famed flexibility as a potential epiphenomenon that can, but need not, emerge from the many ways we go about holding and transforming information. By acquiescing to a pessimism about their explanatory range, we allow cognitive categories to do what they do well; that is, to help marshal, corral, and organize candidate computational structures—and their implementations—without requiring them to pull double-duty as explanatory fulcrums for the higher or universal properties of cognition. Adopting this productive pessimism blunts the demand for natural kinds in a future cognitive science, as it removes the need for our cognitive categories to explain various mythic and unifying properties presumed of cognition. The trade-off earned by reducing the role of these computationally extravagant properties, and the dim epistemic starting position they place us in (cf. Fodor's *First Law*), is likely a more accurate, if messy, depiction of the many, diverse, and often suboptimal ways we go about thinking and behaving in the world.

Acknowledgements This project could not have been realized without the continued support and feedback from friends and colleagues, and a number of helpful audiences from the ESPP in 2015, the SSPP in 2018, the Neural Mechanisms Online Webconference in 2018, and the 2019 Workshop on Natural Kinds and Cognitive Science hosted by York University. I am especially grateful to Lisa Miracchi and the MIRA group at the University of Pennsylvania, to David Rosenthal and the CUNY Cognitive Science Speakers Series, and to Michael Pauen and his group at the Berlin School of Mind and Brain, for allowing me to workshop iterations of this paper. I am also indebted to Matthew Rachar, Tyler Brooke-Wilson, Jesse Prinz, John Greenwood,

[23] As I envision it, this project would build upon and extend a useful heuristic that I've come to half-jokingly term the *second paragraph rule*. Most empirical papers on working memory begin with a formulaic description of working memory as the capacity that enables us to maintain and manipulate information, etc., before citing Baddeley's or Cowan's or another popular model. Afterwards, and sometimes as soon as the second paragraph, the authors make clear the specific iteration of maintenance or manipulation under investigation. Consult Zanto et al. 2011 paper for an excellent example of this heuristic in action, where their focus is, in fact, understanding how impacts to top down activity projected on visual cortices affects selective attention and recognition. This process, under the productive pessimism that I'm advocating, would be a candidate description for how we maintain information. Ideally, much of the work of perusing the literature could be managed using machine learning or natural language processing, quickly yielding a trove of similar candidate descriptions.

Felipe de Brigard, Shaun Nichols, Eric Mandelbaum, and a slew of anonymous reviewers for their invaluable comments on earlier drafts. Finally, I am particularly grateful for the support of some of my earliest mentors, including Carol Seger and especially Whit Schonbein, for persuading me to look critically at working memory and for their immense help throughout this long process.

References

Adams, E.J., A.T. Nguyen, and N. Cowan. 2018. Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. *Language, Speech, and Hearing Services in Schools* 49: 340–355.

Allen, C. 2017. On (not) defining cognition. *Synthese* 194: 4233–4249.

Aristotle & Hamlyn, D. W. (1993). *De Anima: Books II and III*. New York: Oxford University.

Atkinson, R.C. & Shiffrin, R. M. (1971). *The control processes of short term memory* (Report No. 173). Institute for Mathematical Studies in the Social Sciences: Stanford: USA.

Baddeley, A. 1996. Exploring the central executive. *The Quarterly Journal of Experimental Psychology* 49 (A): 5–28.

Baddeley, A. 2007. *Working memory, Thought, and Action*. New York: Oxford University.

Baddeley, A. 2010. Working Memory. *Current Biology* 20: R136–R140.

Baddeley, A., and G. Hitch. 1974. Working memory. *Psychology of learning and motivation* 8: 47–89.

Baddeley, A., and R.H. Logie. 1999. Working Memory: The multiple-component model. In *Models of Working Memory*, ed. A. Miyake and P. Shah. New York: Cambridge University.

Bergson, H.C. (1990). *Matter and Memory* (N.M. Paul and W.S. Palmer, Trans.), Princeton: Zone Books.

Bogdanov, M., and L. Schwabe. 2017. Transcranial stimulation of the dorsolateral prefrontal cortex prevents stress-induced working memory deficits. *Journal of Neuroscience* 36: 1429–1437.

Boyd, R. 1999. Homeostasis, species, and higher taxa. In *Species: New interdisciplinary essays*, ed. R.A. Wilson. Cambridge: MIT Press.

Brandom, R. 2001. *Articulating reasons: An introduction to inferentialism*. Cambridge: Harvard University Press.

Buckner, C. 2015. A property cluster theory of cognition. *Philosophical Psychology* 28: 307–336.

Carruthers, P. 2014. On Central Cognition. *Philosophical Studies* 170: 143–164.

Carruthers, P. 2015. *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford: Oxford University Press.

Cheng, S., and M. Werning. 2016. What is episodic memory if it is a natural kind? *Synthese* 193 (5): 1345–1385.

Cherniak, C. 1986. *Minimal Rationality*. Cambridge: MIT Press.

Christophel, T.B., P.C. Klink, B. Spitzer, P.R. Roelfsema, and J.D. Haynes. 2017. The distributed nature of working memory. *TRENDS in Cognitive Science* 21 (2): 111–124.

Cowan, N. 1988. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin* 104: 163–191.

Cowan, N. 2008. What are the differences between long-term, short-term, and working memory? *Progress in Brain Research* 169: 323–338.

Cowan, N. 2017. The many faces of working memory and short-term storage. *Psychonomic Bulletin and Review* 24: 1158–1170.

Craik, F.I.M., and B.A. Levy. 1976. The concept of primary memory. In *Handbook of Learning and Cognitive Processes, Volume 4*, ed. W.K. Estes, 133–175. New York: John Wiley and Sons.

Craver, C. 2009. Mechanisms as natural kinds. *Philosophical Psychology* 22: 575–594.

D'Esposito, M., B.R. Postle, D. Ballard, and J. Lease. 1999. Maintenance versus manipulation of information held in working memory: An event-related fMRI study. *Brain and Cognition* 41: 66–86.

Dehaene, S. 2014. *Consciousness and the Brain: Deciphering how the Brain Codes our Thoughts*. New York: Penguin.

D'Esposito, M., and B.R. Postle. 1999. The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia* 37: 1303–1315.

Dragoi, G., K.D. Harris, and G. Buzsáki. 2003. Place representation within hippocampal networks is modified by long-term potentiation. *Neuron* 39: 843–853.

Engle, R.W. 2002. Working memory capacity as executive attention. *Current Directions in Psychological Science* 11: 19–23.
Evans, J.S.B., and K.E. Stanovich. 2013. Dual process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8: 223–241.
Evans, G. 1982. *The Varieties of Reference*. Oxford: OUP.
Fernandez, J. 2019. *Memory: A Self-Referential Account*. New York: Oxford University.
Firestone, C., and B.J. Scholl. 2016. Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences* 1: 1–77.
Fodor, J. 1983. *The Modularity of Mind*. Cambridge: MIT Press.
Funahashi, S., C.J. Bruce, and P. Goldman-Rakic. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* 61: 331–349.
Fuster, J.M., and G.E. Alexander. 1971. Neuron activity related to short-term memory. *Science* 173 (3997): 652–654.
Gazzaniga, M.S., R.B. Ivry, and G.R. Mangun. 2009. *Cognitive Neuroscience: The Biology of the Mind*. New York: W.W. Norton.
Gerstner, W., A.K. Kreiter, H. Markram, and A.V.M. Herz. 1997. Neural codes: Firing rates and beyond. *Proceedings of the National Academy of Sciences* 94: 12740–12741.
Gevins, A., M.E. Smith, L. McEvoy, and D. Yu. 1997. High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex* 7: 374–385.
Godfrey-Smith, P. 2016. *Other Minds*. New York: Farrar, Straus and Giroux.
Goldman-Rakic, P.S. 1995. Cellular basis of working memory. *Neuron* 14 (3): 477–485.
Gomez-Lavin, J. (2017) The centered mind: What the science of working memory shows us about the nature of human thought. *Philosophical Psychology 30*(5):685–688.
Hacking, I. 2007. The contingencies of ambiguity. *Analysis* 67: 269–277.
Husserl, E. (1965). *The Phenomenology of Internal Time-Consciousness* (J.S. Churchill, Trans.; M. Heidegger, Ed.), Bloomington, IN: Indiana University Press.
Ikkai, A., and C. Curtis. 2011. Common neural mechanisms supporting spatial working memory, attention and motor intention. *Neuropsychologia* 49: 1428–1434.
James, W. 1890. *The Principles of Psychology*. Vol. 1. New York: Holt.
Jerde, T.A., E.P. Merriam, A.C. Riggall, J.H. Hedges, and C.E. Curtis. 2012. Prioritized maps of space in human frontoparietal cortex. *Journal of Neuroscience 32* (48): 17382–17390.
Jiruska, P., J. Csicsvari, A.D. Powell, et al. 2010. High-frequency network activity, global increase in neuronal activity, and synchrony expansion precede epileptic seizures in vitro. *Journal of Neuroscience* 30: 5690–5701.
Jonides, J., and D.E. Nee. 2006. Brain mechanisms of proactive interference in working memory. *Neuroscience* 139 (1): 181–193.
Khalidi, M.A. 2013. *Natural Categories and Human Kinds*. New York: Cambridge University Press.
Khalidi, M. A. (2015). Natural kinds as nodes in casual networks. *Synthese*, Special issue: Metaphysics and Causation (online).
Klatzky, R. 1975. *Human Memory: Structures and Processes*. San Francisco: W. H. Freeman.
Klein, S.B. 2015. What memory is. *Wiley Interdisciplinary Reviews: Cognitive Science* 6 (1): 1–38.
Korsgaard, C.M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press.
Kubota, K., and H. Niki. 1971. Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology* 34 (3): 337–347.
Ladd, G.T. 1887. *Elements of Physiological Psychology*. London: Longmans, Green & Co..
LaRocque, J.J., J.A. Lewis-Peacock, and B.R. Postle. 2014. Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience* 8 (5): 1–14.
LaRocque, J. J., Eichenbaum, N. S., Starrett, M. J., Rose, N. S., Emrich, S. M., & Postle, B. R. (2015). The short- and long-term fate of memory items retained outside the focus of attention. *Special Issue on Working Memory] Memory & Cognition* 43 (3), 453–468.
Lewis-Peacock, J.A., A.T. Drysdale, and B.R. Postle. 2015. Neural evidence for the flexible control of mental representations. *Cerebral Cortex* 25 (10): 3303–3313.
Lewis-Peacock, J.A., A.T. Drysdale, K. Oberauer, and B.R. Postle. 2012. Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience* 24 (1): 61–79.
McDowell, J. 1996. *Mind and World*. Cambridge: Harvard University Press.
Michaelian, K. 2011. Is memory a natural kind? *Memory Studies* 4 (2): 170–189.
Michaelian, K. 2015. Opening the doors of memory: Is declarative memory a natural kind? *Wiley Interdisciplinary Reviews: Cognitive Science* 6 (6): 475–482.

Miller, G., E. Galanter, and K.H. Pribram. 1960. *Plans and the Structure of Behavior*. New York: Henry Holt and Company.

Miyake, A., and P. Shah. 1999. *Models of Working Memory*. New York: Cambridge University Press.

Newell, A., and H.A. Simon. 1956. *A logic theory machine and a complex information processing system*. Santa Monica, California: The RAND Corporation.

Petrides, M. 2000. The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research* 133: 44–54.

Polonsky, A., R. Blake, J. Braun, and D.J. Heeger. 2000. Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neuroscience* 3: 1153–1159.

Postle, B.R. 2006. Working Memory as an emergent property of the mind and brain. *Neuroscience* 139 (1): 23–28.

Postle, B.R. 2016. Neural bases of the short-term retention of visual information. In *Mechanisms of Sensory Working Memory: Attention and Performance XXV*, ed. P. Jolicoeur, C. Lefebvre, and J. Martinez-Trujillo, 43–58. London: Academic Press.

Postle, B.R., F. Ferrarelli, M. Hamidi, E. Feredoes, Peterson M. Massimini, A. Alexander, and G. Tononi. 2006. Repetitive transcranial magnetic stimulation dissociates working memory manipulation from retention functions in prefrontal, but not posterior parietal, cortex. *Journal of Cognitive Neuroscience* 18: 1712–1722.

Prinz, J. 2012. *The Conscious Brain*. New York: Oxford University Press.

Quilty-Dunn, J. 2019. Is iconic memory iconic? *Philosophy and Phenomenological Research*. https://doi.org/10.1111/phpr.12625.

Ramsey, W. 2017. Must cognition be representational? *Synthese* 194: 4197–4214.

Repovš, G., and A. Baddeley. 2006. The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience* 139: 5–21.

Rottschy, C., R. Langer, I. Dogan, K. Reetz, A.R. Laird, J.B. Schulz, P.T. Fox, and S.B. Eickhoff. 2012. Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage* 60 (1): 830–846.

Shallice, T., and E.K. Warrington. 1970. Independent functioning of verbal memory stores: A neuropsychological study. *The Quarterly journal of experimental psychology 22* (2): 261–273.

Shepard, R.N., and J. Metzler. 1971. Mental rotation of three-dimensional objects. *Science* 171: 701–703.

Sprauge, T.C., E.F. Ester, and J.T. Serences. 2016. Representing latent visual working memory representations in human cortex. *Neuron* 91: 694–707.

Stokes, M.G. 2015. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences* 19 (7): 394–405.

Völter, C. J, Mundry, R., Call J., & Seed, A. M. (2019). Chimpanzees flexible update working memory contents and show susceptibility to distraction in the self-order search task. *Proceedings of the Royal Society*, 286(B), (online).

Wang, K.S., D.V. Smith, and M.R. Delgado. 2016. Using fMRI to study reward processing in humans: Past, present, and future. *Journal of Neurophysiology* 115: 1664–1678.

Zanto, T.P., M.T. Rubens, A. Thangavel, and A. Gazzaley. 2011. Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nature Neuroscience* 14: 656–661.