Computation as the Boundary of the Cognitive

Daniel A. Weiskopf

Abstract: Khalidi identifies cognition with Marrian computation. He further argues that Marrian levels of inquiry should be interpreted ontologically as corresponding to distinct semi-closed causal domains. But this counterintuitively places the causal domain of representations outside of cognition proper. A closer look at Khalidi's account of concepts shows that these allegedly separate Marrian domains are more tightly integrated than he allows. Theories of concepts converge on algorithmic-representational models rather than computational ones. This suggests that we should reject the wholesale identification of cognition with computation.

Khalidi's book is distinctive within recent discussions of cognitive ontology insofar as it puts ontology first, building on his previously established conditions on kindhood. In this discussion, however, I want to consider the *cognitive* part of cognitive ontology. What makes a state, process, capacity, or phenomenon cognitive?

In Khalidi's telling, cognitive phenomena are a subset of psychological phenomena that are defined by exclusion: cognition is "opposed to affective, perceptual, sensory, or experiential aspects of mentation" (p. 3). What this negative definition leaves behind can be gleaned from the book's later chapters: conceptual thought, language use, non-procedural memory, and reasoning heuristics, as well as pathologies of thought like Body Dysmorphic Disorder. In other words, "cognition" is implicitly restricted to what is sometimes called *higher* or *central* cognition.

But Khalidi also offers a theoretically deeper account of cognition that draws on Marr's hierarchy of explanatory levels: "As I conceive it, the domain of cognition lies broadly within what Marr (1982) famously identified as the computational level, as opposed to the algorithmic and implementational levels" (p. 27; cf., p. 236). Later, he states the point slightly differently: "the computational level identifies causal systems that are roughly coextensive with the cognitive

domain." (p. 126). The first description allows that cognition is only one subset of what exists "at the computational level"; the latter suggests that they coincide. The former description is clearly preferable, since all sorts of non-cognitive devices—cash registers and the like—have computational analyses. Moreover, identifying the computational level with higher cognition would disbar vision, Marr's own research domain, from counting as computational.

Still it seems puzzling, bordering on a category error, to identify the domain of the cognitive—something ontological—with Marr's computational level. Marr's levels expressly correspond to distinctive investigative *questions*, or to the pursuit of different forms of description and explanation. They concern ways of framing kinds of inquiry about systems, rather than the ontic constituents of the systems themselves. We can seek to describe the computation that a system is carrying out and conjecture what makes it appropriate given the system's goal or purpose, or we can investigate the representations and algorithms that the system employs, or we can uncover how those algorithms are implemented in the system's physical structure. It is an open question to what degree these three investigative questions correspond to separate domains of properties and entities. Ontology and inquiry have no inherent relationship.

To bridge this gap, Khalidi offers a best-explanation argument for thinking that these methodological distinctions carry ontological import (pp. 27-8). Using the example of vervet alarm calls, he says that a computational account would describe the mapping from the monkey's visual and auditory inputs to their vocal outputs and "explain why these specific predators are the ones that elicit these calls" (p. 27). The explanatory goal is to state why this mapping in particular would be appropriate given the history and environment of the system. He proposes that the success of such explanations turns on the existence of "relatively self-contained causal

processes in the cognitive domain that can be understood somewhat independently" of their implementation in algorithmic-representational processes and neural systems.

So the fact that Marrian strategies of computational analysis can be successfully applied to a system indicates that it is in some sense a real computational device: it has causal properties that correspond with elements of the computational description. This is meant to establish that computation has ontological weight. Moreover, Khalidi says, within this ontological domain "computational or cognitive systems are functionally individuated, since what a system does in solving a problem or performing a task is naturally understood in terms of its function" (p. 28). The tacit argument here seems to be that computational systems are real and functionally individuated, and in this respect they are analogous to cognitive systems. The implied conclusion is that we should take cognitive systems to be a subclass of computational systems, which in turn are a subclass of functionally individuated systems more generally.

Despite its far-reaching implications, this argument goes by surprisingly quickly. The identification of cognition with computation proceeds from a general functionalist premise. But functionalism per se does not single out any one of Marr's levels as the locus of cognition. Algorithms and representations are also standardly defined in schematic functional terms that abstract away from detailed realization properties. For that matter, neural descriptions, too, can be given in terms ranging from highly structure-specific to comparatively noncommittal (as in many network models of brain function). If cognition is uniquely identified with one of these domains—not itself an obvious premise—the algorithmic has at least an equal claim with the computational.

Identifying cognition with computation has other consequences. Most strikingly, it implies that there are no strictly cognitive *processes*, since computational analysis is not a form

of process analysis. In Marr's scheme, computational analysis characterizes a system's inputs, its outputs, and the mapping between the two. Consider stereopsis (Marr 1982, pp. 111-116), in which the inputs are images from each eye ("primal sketches") and the output is the distance to the proximal surface of a focal object. Solving this problem involves locating unambiguous matching points in each image given the intensity changes available within them and then measuring the angular disparity of those points. Computational analysis thus describes the function of stereopsis in the logico-mathematical sense: an input-output mapping paired with a description of the principles by which the system derives outputs from inputs. No processes for executing this are mentioned; that is the province of algorithmic analysis.

In making Marrian levels ontological, questions of representation and process are bracketed off from the domain of cognition. Marr drives home this point in his discussion of transformational grammars: their "operations" are a theorist's ways of describing the function from input string to parsed sentence, not psychologically real processes (1982, p. 357). Since the "objects" in the cognitive domain are idealized and abstracted capacities characterized only by their input-output profile, investigations into how they work is not part of the study of cognition at all.

The same dialectic occurs in contemporary debates over Bayesian models of cognition. Bayesians argue that cognitive systems should be analyzed as executing rational or optimal forms of probabilistic inferences, wherein inputs are assigned probabilities and in combination with the system's prior probabilities generate as output a posterior probability. Computationally, the system is taking the product of the prior and the likelihood, but this idealized inferential pattern can be approximated by any number of different algorithms, each of which predict their

own patterns of performance and breakdown. Computations constrain process models, but don't themselves describe any such processes.

By centering computation as he does, Khalidi aligns himself with a prominent movement in cognitive psychology that advocates theorizing exclusively in computational terms. This movement includes so-called Bayesian fundamentalists, according to whom "[cognitive] theory is cast entirely at the computational level (in the sense of Marr 1982), without recourse to mechanistic (i.e., algorithmic or implementational) levels of explanation" (Jones & Love 2011, p. 175). A similar view is expressed by Columbo & Series (2012): "the methodology adopted is typically performance-oriented, instead of process-oriented… Unlike process-oriented models, performance-oriented models treat their targets as systems that exhibit overall properties. No internal structure is specified within the model" (p. 707). Computational fundamentalists hold that computational theorizing is sufficient on its own for cognitive explanation. Khalidi expresses this view insofar as he thinks that Marrian styles of analysis will correspond, ontologically, to semi-closed domains of causal systems.

However, looking at contexts where the Marrian ontology is put to work reveals the tight interweaving of these domains, as well as the continuing importance of the algorithmic-representational level of analysis. In Khalidi's discussion of concepts (Ch. 2), he proposes splitting the kind *concept* into three subsidiary kinds corresponding to separate computational, algorithmic-representational, and implementing causal structures. He associates each of these with distinct theories, and concludes that the psychology of concepts has been systematically confused about which Marrian levels it is investigating. The result is that it sees illusory conflicts between theories that concern entirely different subject matter. Specifically, Khalidi holds that prototype theorists are advancing an algorithmic-representational account: "it is natural to think

of prototype theory as providing an algorithm for concept activation" (p. 62). Theory theorists are engaged in analyzing concept possession in computational terms: "By contrast, the [theory] theory of concepts is more closely related to Marr's 'computational level.' At the computational level of analysis and explanation, the emphasis is on what concepts enable cognitive agents to achieve and why they possess the concepts that they do." (p. 63). Finally, "[t]he modal theory of concepts is clearly framed in implementational terms" (p. 67). We therefore have kinds belonging to separate ontological domains corresponding to each of these theories, and hence no conflict among them.

I propose an alternative picture of the theories invoked here and the styles of analysis they are paired with. Take the computational domain first. Khalidi associates this with the "theory theory" of concepts, and also links it with specifying what it is to possess a concept. This first linkage is problematic insofar as the theory theory as it is presented is not a computational account at all, since it fails to specify any cognitive capacity and its distinguishing input-output function. If we turn our attention to existing computational accounts of categorization, category learning, and induction, we also find them to be largely uninterested in questions of concept possession. In fact, as in most of psychology, they take criteria of possession for granted. These models focus instead on discerning the rational structure of the task and its optimal solution. They draw on much of the same evidence that informs algorithmic-representational theories, and aim to provide a unifying framework within which representational questions can be pursued.

Consider an example. One well-known phenomenon in category learning is the existence of shifts from one type of representation to another. Learners may begin acquiring concepts by generating abstract prototypes, then shift to a memorized exemplar cluster strategy (Smith & Minda 1998), or even to "mixed" representational forms that are hybrids of or intermediate

between the two. While categorization processes and their associated representations are reasonably well understood, less is known about the mechanisms that drive strategic switching among processes. However, the reasons why such shifts make sense—i.e., are a rational or cogent response to the task and environment—can be explicated within a Bayesian framework.

Griffiths and colleagues (2007, 2008) created a rational model of category learning that aimed to generalize over existing models of representational shifts. Their model treats category learning as the outcome of a hierarchical Dirichlet process (HDP). Objects in the model are assigned to clusters based on the features they share and categories are formed by grouping clusters together. In simple prototype theory, there is only one cluster per category and clusters are never shared (i.e., categories are defined by their prototypes). In exemplar theory, each object gets its own cluster and again there is no sharing of clusters. However, the HDP model leaves it open whether clusters are shared across categories or not. The model has two parameters, α and γ: α controls how many clusters there are for each category, while γ controls how much sharing of clusters there is among categories. So it is possible for there to be several clusters within a single category (if it contains several related prototypes), or for several categories to share a single cluster. The learning rules allow the formation of an unbounded number of new clusters as new objects are encountered, and the parameters jointly determine when cluster formation occurs.

The HDP model can mimic the behavior of prototype and exemplar models when α and γ are set to limiting values of $(0, \infty)$ and (∞,∞). It can also generate mixed behaviors that belong to neither model. Most saliently, it can be configured to add new clusters when the learning data demands it, but not to share these clusters across categories. These settings allow it to produce the best fit to human learning data that demonstrate the prototype-to-exemplar shift. On a

stimulus set containing mostly prototype-conforming items and a single exceptional exemplar, a mixed HDP model will switch strategies midway, allowing it to outperform either solo category learning strategy. This is what Smith and Minda's participants also did. The HDP model explains this in terms of how a rational thinker would learn given certain assumptions about the structure of the environment. Forming a new cluster (and therefore a new category) is warranted given the parameter settings which encode these assumptions. The computational model itself makes no claims about the specific processes and mechanisms involved, nor, crucially, about the form of the learned representations themselves.

Notably, all of this assumes that prototype and exemplar representations are in conflict with each other. The explanatory question centers on how to account for learning phenomena that neither model can capture on its own. Contra Khalidi's claim that concepts are not split into representational subkinds, then, computational analysis takes facts about representational diversity and dynamics as evidence to be accounted for. Moreover, it attempts to give a unifying explanation of why each kind is used under differing circumstances, along with a preliminary characterization of possible hybrid or intermediate representational kinds that can be further investigated using process-based algorithmic analysis and modeling. Computational theorizing about conceptual capacities leads us back to a view on which there *is* a genuine debate among models of concepts, but a debate that is conducted within algorithmic-representational inquiry.

Similarly, far from being confined to implementational matters, modal theories of concepts are explicitly interested in the representational level, and are rich in proposals about conceptual processes such as simulation construction, feature generation, feature verification, etc. While Wu and Barsalou (2009) refer in passing to "perceptual simulation in the brain" in framing their hypothesis, their studies of conceptual combination use no neural measurements at

all, only verbal materials and behavioral outcomes. Their conclusion that perceptual representations are generated in interpreting novel noun compounds is supported by analysis of the kinds of properties that participants produce (e.g., "glass car" elicits information about the car's insides). The same holds for Solomon and Barsalou's (2005) conclusions that perceptual effort affects response time in property verification tasks (e.g., responding yes/no to "does a gorilla have a face?"), as well as for Pecher, Zeelenberg, and Barslaou's (2004) studies of modality switching costs in property verification. All of these support modality-specific theories of conceptual representation without drawing on evidence concerning implementation.

When neural evidence is cited, it is framed as contributing towards algorithmic-level hypotheses. Barsalou (2017) makes this point in his critique of attempts to create a "semantic atlas" of cortex by using neural encoding and decoding techniques paired with multivoxel pattern analysis to map regions corresponding to various lexicalized categories. These methods amount to a form of "neurobehaviorism" in which computationally characterized phenomena are mapped directly onto neural structures without an intermediary account of representations and processes that might explain these implementations (p. 24). Neural data, he maintains, is informative about cognition precisely to the degree that it contributes to building process models with specific commitments to representational formats. Some advocates of modal-specificity may aim to directly characterize conceptual phenomena in neural terms, but this is not an inherent commitment of the theory, only of a particularly reductive form of it.

To conclude, then, functional considerations alone don't mandate locating cognition at any particular one of Marr's "levels." Moreover, there are strong reasons against adopting computational fundamentalism. Defining cognition in a way that eliminates processes renders it impossible to give causal explanations of thought and behavior in cognitive terms. At the same

time, actual computational modeling continually looks “downward” to representational accounts. Computational models of category learning are designed to be consistent with the existence of different representational kinds, to predict the existence of new representational phenomena, and to give a unifying account of when the use of representational kinds is rationally adaptive to environmental demands. Computational theory both draws on and feeds back into algorithmic-representational modeling. Similarly, modal theories of concepts are not just implementational but also aim to contribute directly to representational theorizing. Their representational hypotheses about perceptual simulation in turn predict and unify patterns of neural evidence. Contra what computational fundamentalism suggests, we find a pervasive consilience of top-down and bottom-up theorizing that converges on the algorithmic-representational domain. This suggests that cognitive ontology should not arbitrarily shackle itself solely to the domain of computation.

References

Barsalou, L. W. 2017. What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia, 105*, 18–38.

Colombo, M., & Series, P. 2012. Bayes in the brain—On Bayesian modelling in neuroscience. *British Journal for the Philosophy of Science, 63,* 697–723.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. 2008. Categorization as nonparametric Bayesian density estimation, In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind* (pp. 303–328). Oxford: Oxford University Press.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. 2007. Unifying rational models of categorization via the hierarchical Dirichlet process. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 29.

Jones, M., & Love, B. C. 2011. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*, 169–88.

Marr, D. 1982. *Vision*. Cambridge: MIT Press.

Pecher, D, Zeelenberg, R., & Barsalou, L. W. 2004. Sensorimotor simulations underlie conceptual representations: Modality-specific effects of prior activation. *Psychonomic Bulletin & Review, 11*, 164–67.

Smith, J. D., & Minda, J. P. 1998. Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1411–1430.

Solomon, K. O., & Barsalou, L. W. 2004. Perceptual simulation in property verification. *Memory & Cognition, 32*, 244–59.

Wu, L., & Barsalou, L. W. 2009. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica, 132*, 173–89.