8-2022

A Defense of Algorithmic Homuncularism

Spencer Kinsey

A Defense of Algorithmic Homuncularism

by

Spencer Kinsey

Under the Direction of Daniel Weiskopf, PhD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

in the College of Arts and Sciences

Georgia State University

2022

ABSTRACT

In this thesis, I defend the explanatory force of algorithmic information processing models in cognitive neuroscience. I describe the algorithmic approach to cognitive explanation, its relation to Shea's theory of cognitive representation, and challenges stemming from neuronal population analysis and dimensionality reduction. I then consider competing interpretations of some neuroscientific data that have been central to the debate. I argue in favor of a sequenced computational explanation of the phenomenon, contra Burnston. Finally, I argue that insights from theoretical neuroscience allow us to understand why dimensionality reduction does not militate against localizing distinct contents to distinct components of functioning brain systems.

INDEX WORDS: Cognition, Representation, Computation, Function, Explanation, Reduction

A Defense of Algorithmic Homuncularism

by

Spencer Kinsey

Committee Chair: Daniel Weiskopf

Committee: Andrea Scarantino

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

August 2022

DEDICATION

For Alexander.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dan Weiskopf, for comments and guidance that greatly improved this project, and for his patience and willingness to discuss all of the intricacies of the literature. I would also like to thank my committee member, Andrea Scarantino, whose comments shaped the development of the project during its later stages. Thanks to Mukesh Dhamala and to all of the members of the NeuroPhysics group at Georgia State University for helpful discussion of the material. Thanks to John Bickle and the Deep South Philosophy & Neuroscience Workgroup for hosting a presentation of an earlier version of this work. Finally, thanks to George Wrisley for first inspiring my interest in philosophy.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

One strategy for investigating hypotheses about cognitive function proceeds through the comparison of brain activity measures with algorithmic, or discretely-sequenced, computational models. Shea (2018) has recently proposed a theory of cognitive representation that draws upon this explanatory strategy. According to Shea, we can justifiably assign distinct contents to distinct parts of the brain by determining the unique roles those parts play within algorithmic processes for accomplishing behavioral tasks.

However, some philosophers of mind and brain argue against this approach to cognitive explanation (Anderson, 2014; Burnston, 2021). For instance, Burnston (2021) argues that certain analytic techniques in systems neuroscience, such as dimensionality reduction, suggest that neural systems function to represent complex combinations of information about experimental parameters, and that this undermines algorithmic models of brain function. The central issue of contention is whether algorithmic information processing models can accurately describe the functional properties of brain systems. Burnston answers in the negative, proposing that we reject Shea's approach, which he calls *algorithmic homuncularism* (AH), in favor of an alternative view called *algorithmic coherence* (AC), according to which brain processes may correlate with algorithmic transformations of content even if distinct contents cannot be assigned to distinct brain areas.

In this thesis, I defend the explanatory force of algorithmic information processing models in cognitive neuroscience. In §2, I describe the approach. In §3, I turn to a discussion of the challenges raised by neuronal population analysis, dimensionality reduction, and Burnston's critique of AH. In §4, I respond to these challenges. I first give attention to some data which have been central to the debate (Aoi et al., 2020; Mante et al., 2013). After considering

competing interpretations of the data, I argue that our interpretation of the system is not constrained in the way that Burnston suggests. I then consider the debate in light of developments in theoretical neuroscience that focus on the dimensionality and decodability of neural representations. I argue that the hypotheses advanced by researchers on these empirical fronts allow us to understand why Burnston's critique of AH is problematic.

2 ALGORITHMIC EXPLANATION IN COGNITIVE NEUROSCIENCE

In this section, I describe the practice of developing algorithmic models of cognitive function, and I articulate the explanatory and interpretive leverage that successful models can provide. One way to investigate hypotheses about cognitive function is to compare brain activity measures with algorithmic, or discretely-sequenced, computational models (Love, 2015; Mack et al., 2013). According to some computational and cognitive neuroscientists, interactions between brain areas transform information in a manner necessary for accomplishing behavioral tasks, thereby giving rise to more complex cognitive capacities (Di Carlo et al., 2012; Marr, 2010). For example, activity in some brain area may correspond to the presentation of a stimulus, while activity in another area may carry information about a decision signal, and activity in another may carry directive information about movement. A key aim of the approach is to localize exploitable information to distinct brain regions in order to illuminate the causal structure of thought. An algorithmic information processing model can explain a cognitive phenomenon when there is a plausible mapping between elements of the model and parts of the brain (Kaplan & Craver, 2011; Shea, 2018).

However, Burnston (2021) has recently offered a critique of this approach, which he calls *algorithmic homuncularism* (AH). He writes:

> On this kind of view, distinct physical parts of a system serve as vehicles for distinct contents, and the causal interactions between those vehicles

> implement the content transformations called for by an algorithm… However, I will argue that there are strong reasons to question AH. In particular, certain forms of complexity in neural population responses prevent assigning different contents to spatially and temporally distinct parts of the system. (Burnston, 2021 pp. 1617-1618)

Burnston targets Shea's (2018) theory of cognitive representation, as Shea's theory builds on the central tenets of AH.[1] Here, my primary aim is not to defend Shea's theory of cognitive representation against alternative theories.[2] Instead, the primary contribution of this thesis is to address a cluster of challenges, which I take Burnston (2021) to have made explicit, to the explanatory scope of the computational approach that underpins Shea's theory. Burnston argues against Shea's (2018) version of the *vehicle realism* thesis, according to which a representational description explains when it captures the particular way that a system accomplishes a task function.

I will note three things at the outset that any party to the debate would be remiss not to acknowledge. First, the explanatoriness of any type of cognitive model is a substantively empirical matter: whether a model is explanatory will be constrained both by the structure and content of the model in question and by facts about neural architecture. Second, such models can fail for a variety of reasons: sometimes models should be rejected entirely, but in other cases they may simply need to be refined (Bechtel & Richardson, 2010). Third, and as a corollary of the first point, it is currently an open empirical question whether, or to what extent, brain processes can be decomposed in a way that satisfies reductive desiderata (Shea, 2013).

---

[1] I use the term 'cognitive' to delineate phenomena relevant to the study of mentality, rather than in a narrower sense meant to distinguish it from other elements within an ontology of mind. Moreover, I draw upon Coelho Mollo's (2021) use of the term 'cognitive representation' to refer to the sub-personal representations often studied in cognitive neuroscience, distinguishing them from personal-level states such as propositional attitudes. Shea states that he does not necessarily expect his theory to cover the latter (Shea, 2018).

[2] See Millikan (1984) and Neander (2017) for two prominent alternative teleosemantic theories.

Recapitulating Shea's (2018) theory of cognitive representation falls outside the scope of this thesis, but in order to adequately address Burnston's challenge, it will be necessary to articulate some of the key features of Shea's account. First, I will provide some initial remarks about terminology. For the purposes of this discussion, the term *vehicle* will be used to refer to a concrete, physical part of a system that carries representational content. Parts of the brain can be vehicles, where 'parts' refers broadly to everything from parts of neurons such as dendritic spines, to neurons themselves, to functional assemblies of neurons, and entire brain regions. The term *content* will be used to refer to what a vehicle represents. For example, if brain area MT represents information about motion direction in some experimental context, then MT is a vehicle of content of motion direction information (Burnston, 2016).

Additionally, I will use the term *algorithm* to refer to a series of well-defined steps for achieving some input-output mapping, and the term *computation* to refer to an input-output mapping (Marr, 2010).[3] For example, recognizing a specific person may depend upon sequentially calculating basic visual features, such as edges and orientations, and then higher-level configurational properties, such as facial characteristics, and so on. Often times, more than one algorithm will be suited to a given computation. On Shea's view, identifying the actual algorithm that the brain uses to accomplish a task allows us to determine which contents to assign to which brain regions (Shea, 2018). For instance, multiple algorithms might be suitable for computing the distance between oneself and some object. The idea is that, among the possible algorithms, there is a fact of the matter about which one is playing out within the brain.

[3] At least according to some, algorithmic computation is a form of digital computation, in that it involves manipulating strings of digits (Piccinini & Scarantino, 2011). However, here I assume that not all algorithmic processes are digital. Instead, I take algorithms to be step-wise processes consisting of discrete transitions between internal states, which is compatible both with Shea's view and with the view that neural computation may be a non-digital, or *sui generis* form of computation (Piccinini & Bahar, 2013; Piccinini, 2020). On my view, the brain may implement algorithms even if its processes do not involve digital computations.

According to Shea, vehicles carry content in virtue of bearing *exploitable relations* to the world (Shea, 2018). An exploitable relation is, straightforwardly, a kind of relation that an organism or system can take advantage of when performing a task. Shea describes two types of exploitable relations: correlation and structural correspondence. Correlations can be exploited when they are informative of the probability of world conditions. For example, the firing of a group of neurons in my brain may indicate a greater probability that the driver in front of me has applied the brakes. Internal processes are sensitive to this information, allowing me to quickly deduce that I should apply my brakes. Shea's discussion of structural correspondence is complex. Leaving out some detail, Shea considers a structural correspondence to be a homomorphism, or mapping, between relations on vehicles and relations on what those vehicles represent. For example, Shea cites the well-known study of 'place cell' firing in rat hippocampus as evidence that structural correspondence can obtain in the brain. Place cells are groups of hippocampal cells that fire when animals travel to certain spatial locations (Grieves et al., 2020). Neuroscientists hypothesize that rats can exploit the map-like relations between sequences of place cell firing and spatial routes to determine how to run mazes more efficiently (Shea, 2018).

Shea's (2018) theory also utilizes the notion of a *task function.* A task function is a system output 1) that can occur for some range of inputs and in a variety of contexts, and 2) that the system has been stabilized to produce. To understand the first criterion, it will be helpful to consider the following example. If part of the brain functions to produce information about object class (animate or inanimate, for instance), it should be able to do so regardless of whether a person is browsing in a museum or hiking in the forest. Moreover, it should be tolerant to some variation in input properties, such as shape and size. For example, it should be able to classify both cats and humans as animate, even though there are some obvious differences. The second

criterion pertains to the notion that systems tend to produce certain outputs because doing so in led to positive consequences in the past. For instance, natural selection facilitated the development of internal sensory functions, such as tactile, auditory, and visual functions, because organisms that could not appropriately detect variation in the environment did not survive. However, other learning mechanisms, such as classical conditioning or learning with feedback, can also stabilize function in the relevant sense, and may do so within shorter periods of time.

Shea acknowledges the fact that the two exploitable relations he discusses – correlation and structural correspondence – are both ubiquitous within the brain (Shea, 2018). For example, firing within an area of cortex can spuriously correlate with firing in other brain areas, and with very many external conditions. This gives rise to a substantive question about how we are supposed to determine which correlations are privileged with respect to content ascriptions. Shea introduces the notion of *unmediated explanatory* (UE) information to deal with this problem.[4] Shea defines UE information as follows:

> The UE information carried by a set of components R_i in a system S with task functions F_j is the exploitable correlational information carried by the R_i which plays an unmediated role in explaining, through the R_i implementing an algorithm, S's performance of task functions F_j. (Shea, 2018 p. 84)

Most simply, UE information picks out the content that best captures a component's functional contribution to an algorithmic process that explains a given behavioral capacity. As discussed, one of the problems that this move addresses is the problem of how to determine, among co-extensive correlations, which content to ascribe to a brain region. The central idea is that a particular correlation may be more explanatory of a component's functional contribution to a

[4] Shea also introduces the idea of an unmediated explanatory structural correspondence, but the distinction between UE information and UE structural correspondence will not bear upon this discussion. Therefore, I will refer to UE information as a placeholder for both.

cognitive process than other correlations. For example, within some experimental context, firing within a brain area (say, FFA) might correlate with the following conditions: *an object is present, an animal is present,* and *a face is present.* In order to determine which of these correlations best explains component functioning, we need some way of testing for the region's specific contribution to task performance (Shea, 2018). To draw an example from cognitive neuropsychology, a patient with an injury to the area might be capable of recognizing objects and animals, but incapable of recognizing faces. Careful investigation might shift the balance of evidence in favor of the hypothesis that the region contributes to successful task performance because it transforms inputs so as to make face-specific information available to the cognitive system (Davies, 2010). In that case, face-specific information would be the UE information carried by the area. Accordingly, if damage to the region did not inhibit task performance, that would be a good reason to believe that it did not carry UE information about the task.

The keystone of Shea's view is the linking of UE information with representational content: Shea claims that if a component carries UE information about world condition C, then that component represents C (Shea, 2018). This move allows Shea's theory of representation to appeal to certain influential research strategies prevalent within the mechanistic explanatory framework – namely, decomposition and localization. Decomposition and localization respectively involve analyzing the functioning of an entire system into subfunctions and mapping subfunctions to distinct system components (Bechtel & Richardson, 2010). Bringing this strategy to bear upon the study of behavioral capacities can render algorithmic models of information processing that can then be tested against measures of brain activity (Love, 2015). Shea outlines the approach as follows:

> The first step is to find which computations the subjects could be performing: algorithms that are capable of producing the observed pattern

> of behaviour… The second step is to go into the brain to see which potential algorithm is most consistent with neural activity… When areas show up as potentially representing quantities called for by the algorithm, we check that it is plausible, in terms of neural architecture, that they are computing those quantities in the right sequence (Shea 2018, p. 85)

The strategy thus proceeds by decomposing a cognitive function into subfunctions, determining the kind of UE information that would be necessary for carrying out those subfunctions, and then looking for evidence of where that information is present within the brain. Representations are then identified as components of the brain's information processing mechanisms (Bechtel, 2014).

Researchers have pursued this type of strategy while adopting increasingly sophisticated analytical methods. As a case in point, the following passage from Roskies (2021) captures the manner in which a multivariate model-based fMRI technique called representational similarity analysis (RSA) (Kriegeskorte et al., 2008) can be used to develop and guide hypotheses about algorithmic transformations of content within the brain:

> If RSA reveals a stimulus or task feature that at one stage of processing contributes to differences in the similarity space, but at a next stage that feature appears as an invariant, we can make inferences about the underlying computations and/or intervening representations. For example, in early visual cortex face stimuli do not cluster together in RDMs, but in higher levels of the visual pathway, such as IT, they form a distinct similarity cluster. In a later stage, similarity measures for individual faces do not differ even when the face stimuli are presented in different orientations (Guntupalli et al. 2017), suggesting that identity is computed between these stages of the visual hierarchy. By allowing us to probe which kinds of stimuli or behaviors result in invariances in the similarity matrix, and to look for the emergence of such invariants, we can infer where and when in the processing hierarchy certain higher-order properties are computed/extracted from the signal. (Roskies, 2021 p. 5929)

This support is tentative, and the RSA method makes assumptions, some of which move beyond the assumptions of more traditional neuroimaging analyses (Gessell et al., 2021; Roskies, 2021; Weiskopf, 2021). On my view, hypotheses about cognitive functioning and representation will

ultimately need to be underwritten by interventions.[5] However, the foregoing practices compel further investigation of the prospect of localizing representations in the brain. Where possible, success will translate into an increase in the interpretability of brain dynamics (Kriegeskorte & Diedrichsen, 2019; Shea, 2013).

[5] See Woodward (2003) for an account of interventions of the relevant sort.

3 POPULATION ANALYSIS AND DIMENSIONALITY REDUCTION

In this section, I focus on a cluster of objections that critics have raised to the idea of localizing cognitive function in this manner. Some critics have expressed doubts about the usefulness of decomposition and localization for explaining cognitive processes (Silberstein & Chemero, 2013). Along similar lines, other critics have noted that entire brain networks can carry information about multiple experimental parameters, or conditions, and they contend that this suggests that the representations of those parameters are not localizable to distinct parts of active cortical systems (Anderson, 2014; Burnston, 2021). Burnston (2021) has recently raised an argument to this effect that draws upon machine learning tools often used to reduce the dimensionality, or to provide useful mathematical summaries of, the variance in complex neuroscientific data. Burnston argues that these techniques militate against localizing distinct informational contents to distinct parts of the brain. Moreover, he suggests that we should all but abandon the aspiration of explaining cognitive processes algorithmically – an aspiration that has been a long-standing goal of cognitive science (Weiskopf, 2021).

To illustrate the concerns that Burnston voices, it will be helpful to describe the groundwork of a certain type approach to neuroscientific data analysis that has been increasingly discussed within the philosophical literature – the 'Hopfieldian' approach, or population doctrine (Barack & Krakauer, 2021; Ebitz & Hayden, 2021). The population doctrine maintains an explanatory focus on patterns of activity distributed across entire neural populations, rather than on the distinct activities of individual population constituents, such as subpopulations of neurons. For example, Willett et al. (2020) pursued this approach when investigating how information about limb movement is reflected within the 'hand knob' region of premotor cortex. They discovered that movements of multiple parts of the body could be decoded, in the sense of being

reliably readout by a pattern classifier, from distributed patterns of population activity. The researchers took this result to undermine the formerly prominent motor homunculus model of movement processing that assigned responsibility for controlling distinct parts of the body to distinct parts of motor cortex. Thus, by focusing upon distributed patterns of activation in the target system, the researchers claim to have made headway in explaining how the motor system functions to track and produce movement.

Willett et al.'s (2020) study is exemplary of the population doctrine. The approach is characteristically distinguished by 1) a focus on distributed patterns of system activity and 2) the development of tools and models that purport to capture the global dynamics of the system. Researchers have adopted the approach in the study of motor functioning (Vyas et al., 2020), perceptual decision-making (Mante et al., 2013), and in the synthetic neurophysiology of recurrent networks (Fanthomme & Monasson, 2021).

Philosophers such as Burnston (2021) and Anderson (2014) argue that this population-focused approach to cognitive explanation militates against the localization of representations in brain systems. For instance, Burnston claims that population-focused researchers often deploy dimensionality reduction techniques, which provide useful mathematical summaries of how population activity varies with task conditions, and that these techniques undermine algorithmic explanation because they suggest that neural systems function to represent information in a distributed, rather than localized, manner. In particular, Burnston argues that these methods undermine Shea's (2018) approach to cognitive representation because they call into question two key theses that underwrite his view – *injective mapping* and *causal isomorphism* (Burnston, 2021). Burnston claims that injective mapping requires distinct contents to be mapped onto spatially and functionally distinct vehicles, as opposed to being mapped to shared,

multifunctional vehicles. He also claims that causal isomorphism requires there to be well-defined, discrete causal transitions between vehicles that mirror those posited within an algorithm.

However, Burnston argues that the techniques that systems-focused researchers use to reduce the dimensionality of complex data render algorithmic explanations of neural functioning untenable (2021). To clarify the nature of Burnston's challenge, more must be said about dimensionality reduction. Dimensionality reduction tools such as principal components analysis (PCA) and linear discriminant analysis (LDA) are often deployed in order to find dimensions of variation that best capture the structure of clouds of data points.[6] In neuroscientific applications, these dimensions are most often viewed as weighted patterns of neural activation which, when combined, yield back the unprocessed data. As a simple illustration, consider *Figure 1*, from Cunningham & Yu (2014).

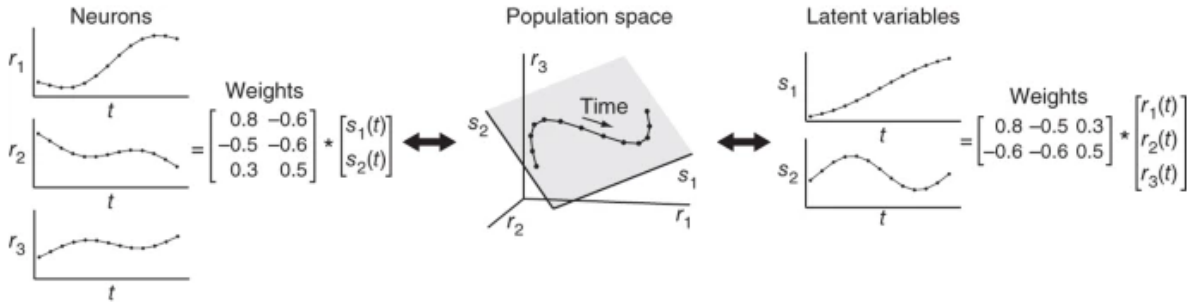

Figure 1 (from Cunningham & Yu, 2014)

On the left, time series data recorded from individual neurons trace out different activity trajectories. At center, the activity of the three neurons is transposed into a three-dimensional state space where the population trajectory can be modeled as a lower (two) dimensional

[6] For a canonical review, see Cunningham & Yu, 2014. The details distinguishing PCA from LDA and other forms of dimensionality reduction are not relevant for the purposes of this discussion.

trajectory.[7] On the right, it is shown that this lower dimensional trajectory can be decomposed into two separate weighted basis, or constituent, activation functions that capture much of the variance in the population trajectory. Because these weighted basis functions are linearly superposed to re-obtain population trajectories, they are often conceived as 'latent variables' inherent in the activity of populations. For this reason, Burnston argues that dimensionality reduction techniques can be useful for simplifying the complex behaviors of neural systems which resist an obvious functional decomposition, or which exhibit mixed-selectivity, in the sense of being activated by multiple distinct experimental task parameters (Burnston, 2021).

The upshot, according to Burnston (2021), is that these techniques do not support dividing neural systems into distinct vehicles that represent distinct contents, and they thus undermine the possibility of achieving an injective mapping between the steps of algorithms and the components of brain systems. Additionally, Burnston argues that these techniques undermine causal isomorphism, as causal isomorphism requires contents to be transformed into other contents during subsequent processing stages. He argues that, if there is overlap between the vehicles that carry the represented contents, then their interactions do not amount to the transformations between contents called for by an algorithm, and causal isomorphism does not obtain.

Burnston (2021) locates a particular difficulty for AH in the fact that dimensions of variation are often drawn from the activity of the entire system targeted by an analysis, rather than from the activities of subpopulations. By drawing our attention to this, he intends to demonstrate that these analyses privilege global functional properties, and therefore tell against Shea's view. Consider the following claims:

[7] A state space is, most simply, a grid-like theoretical space meant to capture all of the possible states that a system might be in.

> …representations are realized as trajectories in the state space of a whole population, rather than as isolated to distinct parts of that population. Hence, injective mapping is false. (Burnston, 2021 p. 1620)

> What is important about this, for current purposes, is that any measure of representation of a particular variable is taken to be constituted by the whole population, rather than by distinct parts within that population. If so, then the system cannot be divided in the way AH recommends. (Burnston, 2021 p. 1624-1625)

This claim exemplifies a distinctive critique of computational and algorithmic explanation in cognitive neuroscience based on dimensionality reduction. On this critique, representations aren't localizable components of a functioning cognitive system, but are instead constitutive patterns or modes of activation of the system, and the graded nature of pattern evolution can prevent us from ascribing distinct contents to distinct stages of activity.[8] Consider Willett et al.'s (2020) analysis of the dynamics within the 'hand knob' region in light of this critique. The researchers concluded that, rather than being confined to functionally distinct subpopulations, information about movement related to various parts of the body was reflected within principal components drawn from the entire area. Furthermore, they found that information about multiple limb movements was reflected within the activity of the system at once – an idea that the researchers refer to as "compositional coding" (Willett et al., 2020 p. 396). With respect to this case, it seems Burnston would argue that limb movement representations are entangled in the activity of the whole population.

Burnston (2021) cites several studies to support his view. In one, Hunt et al. (2015) performed an experiment in which macaque monkeys were trained to choose between pictures that were associated with distinct magnitudes of reward. The researchers performed PCA on local field potential time series data gathered from several prefrontal cortex (PFC) subareas. The

[8] Despite the novelty of the machine learning techniques cited by Burnston, a version of the critique was anticipated by connectionist thinkers, such as Smolensky (1988) and van Gelder (1992).

PCA revealed that the second principal component was correlated with the speed of activity ramp-up and had higher weights on high value trials (Burnston, 2021; Hunt et al., 2015). Hunt et al. took these results to suggest that the calculation of chosen value was an emergent property of population dynamics, and Burnston claims that this suggests that a representation of choice cannot be injectively mapped to a subpopulation or a temporal stage of system activity.

However, Burnston's (2021) critique draws much of its force from a study conducted by Mante et al. (2013). Mante et al. attempted to analyze the dynamics within the macaque PFC during a perceptual decision-making task. In the task, macaques were trained to respond selectively to a dot field stimulus based on a context cue. The field varied both in terms of predominant color (red or green) and in terms of the predominant direction of motion (left or right). The researchers presented the monkeys with a context cue that indicated whether they should select for predominant direction of motion or predominant color. The animals were trained to respond by saccading either to the left or to the right. Stimulus aspects were mixed in a way that was difficult to interpret at the level of individual neuron activations, but PCA revealed that motion, color, context, and choice information were all distinguishable at levels of activation in the population (Mante et al., 2013). Burnston argues that this suggests that the functional properties of the system aren't explainable via Shea's framework, because the principal components corresponding to each task parameter were all weighted positively for the entire population over the duration of the task (Burnston, 2021).

Notably, Shea attempts to apply his account to the Mante et al. (2013) study by modeling the context as an input, various combinations of predominant color and predominant motion direction as a second input, and a processing step which converts these two inputs into an action cue (Shea, 2018). However, Burnston argues that Shea's model is incompatible with Mante et

al.'s (2013) analysis because the PCA results suggest that any representation of an action choice is present throughout the neurons in the population at each stage of activity (Burnston, 2021). Moreover, he argues that this representation is superimposed upon representations of motion, color, and context, suggesting that the entire population acts as a vehicle for all of those contents as it moves throughout a globally-defined state space. Ultimately, Burnston claims that his analysis deflates an algorithmic or computational explanation of the system's behavior, entailing a kind of arbitrariness that refutes Shea's version of vehicle realism.

Burnston (2021) proceeds to offer an alternative view of processing he calls *algorithmic coherence* (AC), according to which brain processes can be correlated with algorithmic transformations of content without being divisible into spatially and temporally discrete representations. For example, where Shea (2018) would speak of the causal processes occurring between functionally distinct subpopulations in terms of implementing the transformations of content called for by an algorithm, Burnston (2021) would say that many neural systems only appear to implement algorithms, but the distinct contents that are supposed to be mapped to the discretely-sequenced steps of an algorithm may be present during each stage of neural activity.

Burnston (2021) considers several objections to his arguments, the predominant one being that expanding our analyses to include other brain regions might reveal that algorithmic models capture brain functioning at larger scales, such as those typically studied by fMRI researchers. For example, perhaps if the Mante et al. (2013) analysis had sampled data from other brain regions, such as the visual cortices, the idea is that we would be able to conceive of the causal transitions between all of the sampled brain regions in terms of implementing the transformations called for by an algorithm. However, Burnston (2021) argues first that this would lessen the explanatory import of AH because it would render activity within brain regions

opaque, and second that we have good evidence that even inter-regional brain networks reflect complex combinations of information in a way that precludes algorithmic explanation. He concludes by proposing an alternative conception of vehicle realism, according to which neural systems can have distinct vehicles representing distinct conditions just in case those conditions have distinct effects on system dynamics, and the dynamics fulfill a task function.

4 A DEFENSE OF ALGORITHMIC HOMUNCULARISM

In this section, I defend the prospects of algorithmic explanation in cognitive neuroscience in light of Burnston's critique. First, I consider a follow-up study of the Mante et al. (2013) data (Aoi et al. 2020), arguing that our interpretation of the system is not constrained in the way that Burnston (2021) suggests. I then turn to a discussion of research in theoretical neuroscience that focuses on the functional dimensionality and decodability of neural activity. I argue that the hypotheses advanced by these researchers allow us to understand why Burnston's critique of AH is problematic.

Recall that Burnston suggests that the primary issue of debate is related to the locality vs. globality of system functioning (Burnston, 2021). He argues that the prefrontal system targeted by Mante et al. (2013) epitomizes a non-decomposable system, in the sense that there is no way to divide the system into spatially and functionally distinct components or into temporally distinct processing stages. He therefore takes its process to be opaque to algorithmic analysis (Burnston, 2021).

However, I will argue that a follow-up study on this data gives us reason to doubt Burnston's interpretation of this system. Aoi et al. (2020) refined our understanding of this system's properties by performing a type of dimensionality reduction that allowed them to analyze fluctuations in how components captured variance over the course of the task. In agreement with Mante et al. (2013), the researchers found that each neuron carried information about every task variable due to the population's "broad tuning characteristics" (Aoi et al., 2020 p. 1410). However, these techniques also revealed that selectivity for task parameters changed for many neurons during the experiment. And, most critically, the researchers discovered that the activity of the population fell into discrete dynamical regimes, first mapping linearly onto an

"early" principal component for each variable, then entering a rotational phase in which it began mapping onto "middle" and "late" principal components that were intended to capture changes in selectivity – cf. *Figure 2* (2020 p. 1415).

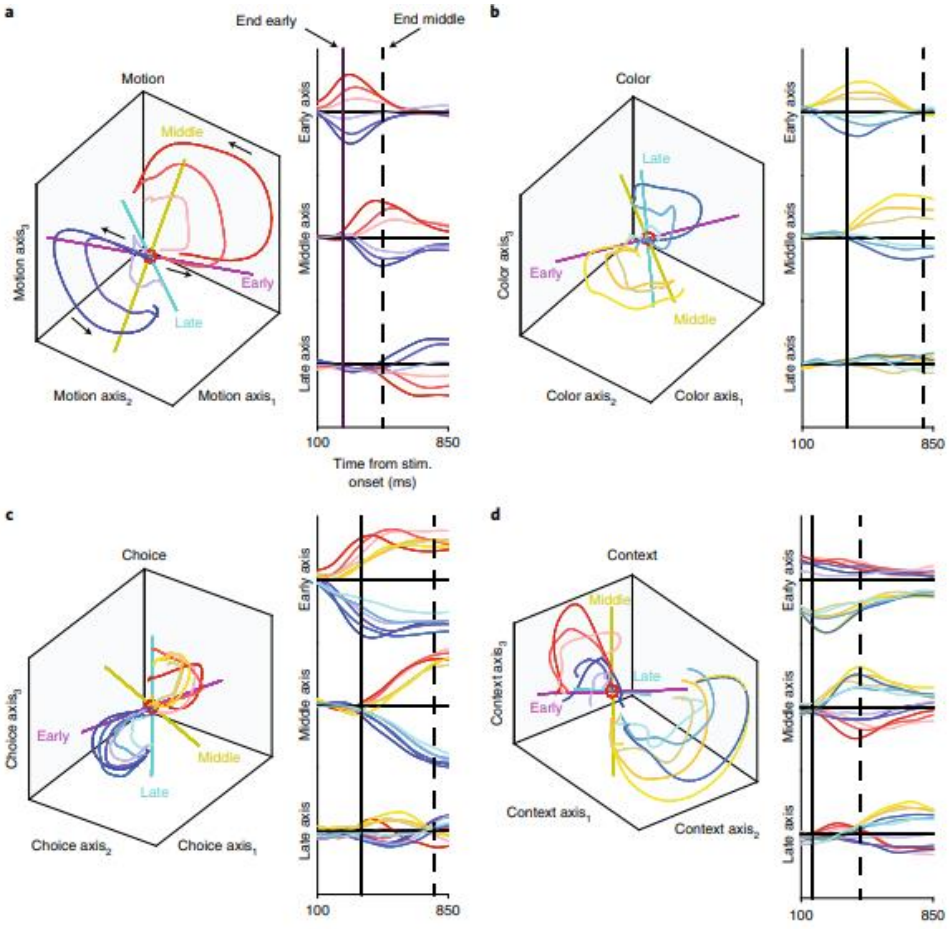

Figure 2 (from Aoi et al., 2020)

Note, from *Figure 2*, how task-related activity first mapped onto the initially explanatory dimension of variation for each parameter, and then began rotating onto the other (middle and late) dimensions. According to Burnston (2021), any representation of choice is the by-product of an ongoing competition between patterns of activation and, as such, he argues that the representation of choice is superimposed on representations of motion and color over the duration of the task. He argues:

> What the temporal division requires is that there be a time at which one content is tokened, but the second content not tokened, and then a specific process that brings about that second content. But in the Mante et al. case, the task axes are constant features of the population response—every location in the PC space describes some place along the task axes. So, there is no temporal stage at which one content is tokened and the others aren't. Nor are there clear phases of transition between some represented contents and others. (Burnston, 2021 p. 1632)

However, Aoi et al.'s (2020) results open up the hypothesis space in a way that casts doubt on this functional interpretation. Specifically, consider the manner in which the researchers compare the transition from the linear to rotational dynamical epochs with the timing of near-peak choice discriminability:

> The temporal separation of the early/linear and rotational subspaces suggests that these are subspaces within which distinct computations are evolving or have independent sets of downstream targets… the present analysis indicates the possibility that the early epochs are concomitant with the temporal window that decision-making is performed. For example, the timing of transition between early and middle epochs is consistent with the timing of accurate decoding of the animals' decisions from single pseudo-trials… This evidence suggests that the transition from linear to rotational dynamics is a correlate of decision commitment. (Aoi et al., 2020 p. 1418)

In this passage, Aoi et al. advance a clear hypothesis about dividing system activity into distinct psychofunctional stages. Moreover, that hypothesis is guided by considerations about downstream decodability. Decodability does not entail representation, but there is some consensus that downstream neural structures need to implement a decoding process at least as complex as that of a linear classifier to *exploit* information transmitted from upstream structures, so considerations about decodability are often taken to constrain hypotheses about the representational properties of neural activity (Kriegeskorte & Diedrichsen, 2019). What is perhaps most notable about the later analysis is the fact that the researchers advance a sequence-focused perspective despite the presence of an informational mixture detected at the level of the

PCA over the duration of the task. This suggests that the researchers believe that the consistently positive component weightings are compatible with the occurrence of a behaviorally relevant functional transition.

At this point in this discussion, it must be re-emphasized that the functional significance of the dynamical transition in question should be investigated with methods that can reveal its ground truth relation to behavior. However, the subsequent analysis provides tentative support for something more like Shea's (2018) proposal. That is, while the system first appears to behave as an evolving structure with several dimensions of variation, it subsequently transitions into some type of representation that can be exploited in order to carry out the task. More conservatively, Aoi et al.'s (2020) results should make us skeptical of the idea that this system's representational properties remain static over the duration of the task, calling into question Burnston's claim that temporal division would be arbitrary for this system.

An objection should be addressed. It might be objected that even if a processing transition occurs, we still do not get a genuine homuncular analysis, which requires mapping distinct subfunctions to *spatially* distinct parts of the system. Burnston (2021) takes this line. For instance, he briefly considers the idea that content transitions might be trackable through the temporal dynamics of neural populations, but he argues that this is insufficient to ground AH, writing that "AH implies both *spatial* and *temporal* divisions between distinct stages of processing" (Burnston, 2021 p. 1627, his emphasis). However, I consider this objection to be something of a red herring: if a behaviorally relevant functional transition occurs, then that undermines Burnston's claim that the representational properties of this system remain static over the course of the task, regardless of whether those properties can be mapped to spatially distinct parts of the system. The current argument is thus meant to undermine Burnston's claim

that the PCA results suggest that a functional division of this system would necessarily be arbitrary, which would deflate computational explanation and refute Shea's version of vehicle realism.

Alternatively, Burnston might accept the idea that some type of functional transition occurs, but object that the causal isomorphism requirement, which maintains that distinct stages of activity must reflect distinct contents, is not met for this system because the PCA results indicate that the population is selective for every task parameter throughout the duration of the task. While this objection requires careful handling, it can be addressed. The notion of UE information can be operationalized to curtail this objection. If we accept Shea's framework, then the fact that a system is selective for each task parameter over the course of a task does not entail that the system *represents* information about all of those parameters, in the sense of carrying UE information about them, during every stage of activity. That is because UE information must be both in an *explicit* format, in the sense of being available for use to the cognitive system (Di Carlo et al., 2012; Diedrichsen & Kriegeskorte, 2017; Shea, 2007), and explanatory of the component's functional contribution to the cognitive process in question. The sense of explicitness that I draw upon here refers to the notion of directness of extraction from the system (Clark, 1992). Consider the fact that all of the information needed to visually distinguish between categories of objects is present at the retina (Di Carlo et al., 2012; Diedrichsen & Kriegeskorte, 2017). Despite this, category information is not formatted at the retina in a way that facilitates explicit transmission of that content – operations that make category information explicit are performed downstream within the visual system. With regard to the Aoi et al. (2020) study, my contention is that, during the earliest stages of activity, decision information is at best implicit in

the system, but it becomes explicitly tokened as a system output at the time identified by the researchers.

To clarify and to elaborate on why Burnston's critique is problematic, I will turn to developments in theoretical neuroscience that focus on the dimensionality and decodability of neural activity. Researchers are beginning to realize that understanding the functional dimensionality of a brain region is key to understanding how activity within that region can be exploited by an organism when performing a task (Ahlheim & Love, 2018; Badre et al., 2021; Fusi et al., 2016). Lower functional dimensionality is thought to be helpful for generalizing to new samples of the same type, while higher dimensionality allows for flexible context-sensitive distinctions between different sets of information – cf. *Figure 3* (Fusi et al., 2016). Panels (a) and (b) depict the firing rates of the three neurons (f_1-f_3) and their responses under four different experimental task conditions. Note that low dimensionality in (a) leads to the inability of a linear classifier to draw a hyperplane through the four conditions and discriminate between them. Alternatively, higher dimensionality in (b) allows a hyperplane to be drawn between any of the four task conditions. Panels (c) and (d) depict activity vectors in two distinct conditions (f_1 and f_2). These panels show that higher dimensionality can impede generalization efficiency.

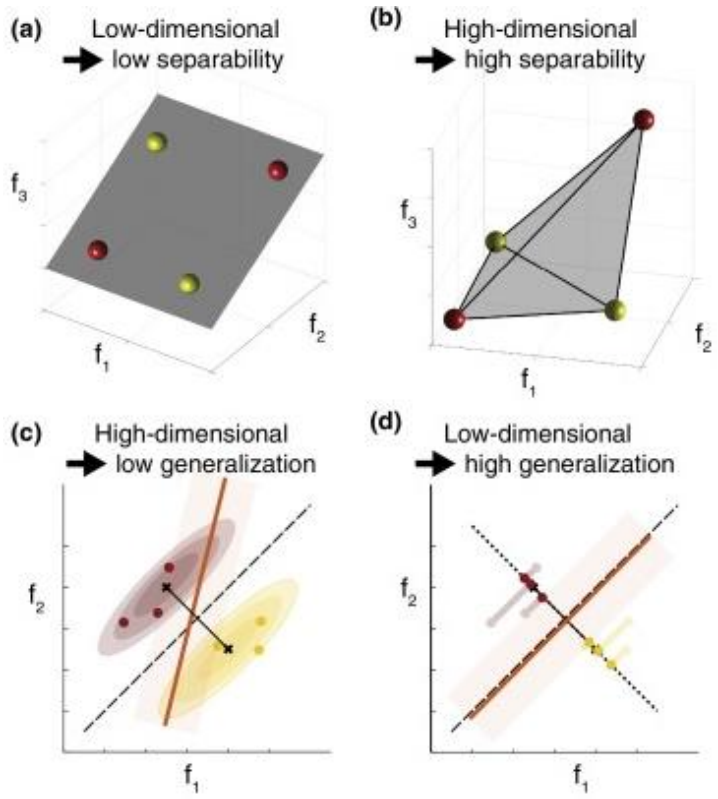

Figure 3 (from Fusi et al., 2016)

For example, consider a set of neurons purely selective for the object category *pyramid.* The firing of these neurons could encode object category in a way that prevents variation in orientation, color, and so on from affecting the output. This type of invariance computation therefore allows for the same classification response regardless of nuisance variation (Fusi et al., 2016), and it is thought to be important for the ability of any intelligent agent to break the 'curse of dimensionality', according to which there are simply too many possible combinations of world-involving features to store them all in memory (Poldrack, 2021).

However, it is also thought that many brain regions exhibit a nonlinear form of mixed selectivity useful for generating large numbers of context-dependent responses (Fusi et al., 2016; Rigotti et al., 2013). Several of the prefrontal cortical areas mentioned previously in this discussion are among such brain regions. These areas are thought to project information into a

high dimensional space which allows for a downstream readout to implement any number of distinctions between task parameters or combinations thereof. This capacity is thought to be especially important for the execution of higher cognitive functions. For example, Rigotti et al. (2013) argue that neurons that exhibit non-linear mixed selectivity for multiple task parameters often exhibit correspondingly higher functional dimensionality in a way that allows for flexible and adaptive context-sensitive behaviors, and that collapses of dimensionality are associated with behavioral errors on task trials. These researchers also note, however, that populations of neurons that exhibit certain forms of mixed selectivity do not exhibit high dimensional responses. In particular, populations that exhibit purely linear mixed selectivity are not thought to exhibit high dimensional representational capacities. This correspondingly constrains hypotheses about the potential roles that they can play within the broader cognitive architecture (Fusi et al., 2016).

The fact that these kinds of considerations are taken to constrain hypotheses about representation leads to a novel argument against Burnston's critique of AH. According to the theoretical neuroscientists, decodability does constrain whether UE information can be ascribed to various stages of activity, and this impugns the idea that ascriptions of content can be made without regard to the *format* in which information is present and available for use. I hypothesize that the process occurring within the system studied by Mante et al. (2013) and Aoi et al. (2020) is one in which the system accumulates and binds evidence into an explicit format so as to facilitate a readout of the appropriate decision in each context. Thinking of system activity in these terms requires qualitatively distinguishing stages of activity. Namely, it requires distinguishing between activity epochs in which the information in the system is not present in a format that facilitates the correct readout, and epochs in which the information is present in such a format. This is incompatible with the proposal that Burnston advances, according to which the

decision representation is a state space trajectory which can be read off of the PCA even during the earliest stages of activity.

Considerations about functional dimensionality and decodability also cast doubt on Burnston's contention that large-scale brain networks will turn out to be cognitively non-decomposable. Recall from §3 that, while considering whether AH might obtain over larger scales within the brain, such as those studied by fMRI researchers, Burnston argues that the best explanations of these processes may also align with a pattern competition model (2021). He writes:

> …it may be that different areas of the brain implement competitions within distinct reference frames—so the competition mediated by the OFC is in a value-frame, whereas dlPFC competitions might take place within a spatio-temporal reference frame. Put differently, it might be that if we go up a level we may get more of the same, now with brain networks representing complex combinations of information in a way that evolves dynamically towards a choice (Cisek and Thura 2018; cf. Anderson 2014; Burnston, forthcoming; Stanley et al. 2019). If so, then AH will not describe between-area processing either. (Burnston, 2021 p. 1634)

Following this, Burnston argues that even if representations can be approximated to distinct brain regions, transitions between contents may not be clear (Burnston, 2021). However, my response to this argument is twofold. First, I contend that even if brain networks come to reflect combinations of information about task parameters, that does not entail that the brain regions involved are all playing the same functional roles, in the sense of transforming information in the same way, or that the processes are most appropriately described as global pattern competitions. To see why, consider how hypotheses about dimensionality and decodability are taken to constrain the kinds of contents we can ascribe to regional activity. If these theoretical tools provide reliable insight into how the information within a region can be *used*, then there is no plausible sense in which representations are state space trajectories of brain networks in the way

that Burnston seems to suggest either. Dimensionality reduction may suggest that information about a particular task parameter is present within several brain regions, but if those regions are not all functioning to make that information explicitly available to the cognitive system, then it is implausible to think that we should ascribe the content to all of those areas. In some instances, shared or mutual information may be functional, while in others it may be non-functional or epiphenomenal. Moreover, distinct brain areas may perform unique computations even when a large amount of information is shared between them (Shehzad & McCarthy, 2018).

With respect to the idea that transitions between contents may not be clear even if we expand our analyses to brain networks, my response is that this an empirical matter. On my view, the type of considerations that I have discussed above should be coupled with causal interventions in order to determine how we should ascribe contents to brain areas. If a particular algorithmic model fails to accurately describe neural processing, we should not hasten to reject algorithmic modeling simpliciter – in some cases it may be that we haven't identified the correct candidate algorithm (Piccinini & Craver, 2011). Furthermore, algorithmic models can remain neutral with respect to how the causal transitions between contents occur, as long as vehicles corresponding to distinct contents in the model can be identified (Shea, 2018).

Finally, an important upshot of the discussion is that we should reject the presumption that we will be able to fully understand the structure of cognitive processing without considering how components are situated within the broader neural architecture. Notably, many of Burnston's arguments against AH depend upon restricting the aperture of analysis in this way. For instance, with regard to the Hunt et al. (2015) study, Burnston argues that "chosen value is not an outcome of a sequence of processing stages that begin with other distinct populations and end in the choice" (Burnston, 2021 p. 1627). However, it is plain that as soon as we expand the

analytical lens to include the activities of other brain areas, including sensory regions, Burnston's assertion here seems much less plausible – the inputs coming into the system must be coming from somewhere.

5 CONCLUSION

I have argued that algorithmic homuncularism remains a viable explanatory strategy in cognitive neuroscience. When brain activity reflects combinations of information about experimental task parameters, that may provide useful constraints on the development of algorithmic information processing models. However, Burnston's critique does not militate against the prospect of localizing cognitive representations in functioning brain systems.

REFERENCES

Ahlheim, C. & Love, B. C. (2018). Estimating the functional dimensionality of neural representations. *Neuroimage, 179,* 51-62. https://doi.org/10.1016/j.neuroimage.2018.06.015

Akam, T. & Kullmann, D. M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience, 15*(2), 111-122. https://doi.org/10.1038/nrn3668

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Aoi, M. C., Mante, V., & Pillow, J. W. (2020). Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature Neuroscience, 23***,** 1410-1420. https://doi.org/10.1038/s41593-020-0696-5

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2020). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences, 38,* 20-28. https://doi.org/10.1016/j.cobeha.2020.07.002

Barack, D. L. & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience, 22,* 359-371. https://doi.org/10.1038/s41583-021-00448-6

Bechtel, W. (2012). Referring to localized cognitive operations in parts of dynamically active brains. In R. Athanassios & P. Machamer (Eds.), *Perception, realism, and the problem of reference* (pp. 262-284). Cambridge: Cambridge University Press.

Bechtel, W. (2016). Investigating neural representations: the tale of place cells. *Synthese, 193,* 1287-1321. https://doi.org/10.1007/s11229-014-0480-8

Bechtel, W. & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research.* Cambridge, MA: MIT Press.

Burnston, D. C. (2016). Computational neuroscience and localized neural function. *Synthese, 193*(12), 3741–3762.

Burnston, D. C. (2021). Contents, vehicles, and complex data analysis in neuroscience. *Synthese, 199,* 1617-1639. https://doi.org/10.1007/s11229-020-02831-9

Clark, A. (1992). The presence of a symbol. In J. Haugeland (Ed.), *Mind design II: Philosophy, psychology, artificial intelligence* (pp. 377-393). Cambridge, MA: MIT Press.

Coelho Mollo, D. (2021) Why go for a computation-based approach to cognitive representation. *Synthese, 199,* 6875–6895. https://doi.org/10.1007/s11229-021-03097-5

Cunningham, J. & Yu, B. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience, 17*, 1500–1509. https://doi.org/10.1038/nn.3776

Davies, M. (2010). Double dissociation: Understanding its role in cognitive neuropsychology. *Mind & Language, 25*, 500-540. https://doi.org/10.1111/j.1468-0017.2010.01399.x

Di Carlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415-434. https://doi.org/10.1016/j.neuron.2012.01.010

Diedrichsen, J. & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology, 13*(4), 1-33. https://doi.org/10.1371/journal.pcbi.1005508

Ebitz, R. B. & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron, 109*(19), 3055-3068. https://doi.org/10.1016/j.neuron.2021.07.011

Fanthomme, A. & Monasson, R. (2021). Low-dimensional manifolds support multiplexed integrations in recurrent neural networks. *Neural Computation, 33*(4), 1063–1112. https://doi.org/10.1162/neco_a_01366

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology, 37*, 66-74. https://doi.org/10.1016/j.conb.2016.01.010

Gessell, B., Geib, B., & De Brigard, F. (2021). Multivariate pattern analysis and the search for neural representations. *Synthese, 199*, 12869-12889. https://doi.org/10.1007/s11229-021-03358-3

Grieves, R. M., Jedidi-Ayoub, S., Mishchanchuk, K., Liu, A., Renaudineau, S., & Jeffery, K. J. (2020). The place-cell representation of volumetric space in rats. *Nature Communications, 11*(789), 1-13. https://doi.org/10.1038/s41467-020-14611-7

Hunt, L. T., Behrens, T. E., Hosokawa, T., Wallis, J. D., & Kennerley, S. W. (2015). Capturing the temporal evolution of choice across prefrontal cortex. *eLife, 4*.

Kaplan, D. & Craver, C. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science, 78*(4), 601-627. https://doi.org/10.1086/661755

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*(4), 1-28. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N. & Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual Review of Neuroscience, 42*(1), 407-432. https://doi.org/10.1146/annurev-neuro-080317-061906

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science, 7*(2), 230-242. https://doi.org/10.1111/tops.12131

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current biology, 23*(20), 2023–2027. https://doi.org/10.1016/j.cub.2013.08.035

Mante, V., Sussillo, D., Shenoy, K., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature, 503***,** 78–84. https://doi.org/10.1038/nature12742

Marr, D. (2010). *Vision.* Cambridge, MA: MIT Press.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, MA: MIT Press.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics.* Cambridge, MA: MIT Press.

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition.* Oxford: Oxford University Press.

Piccinini, G. & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science, 37,* 453-488. https://doi.org/10.1111/cogs.12012

Piccinini, G. & Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese, 183,* 283–311. https://doi.org/10.1007/s11229-011-9898-4

Piccinini, G. & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics, 37*(1), 1-38. https://doi.org/10.1007/s10867-010-9195-3

Poldrack, R. A. (2021). The physics of representation. *Synthese, 199,* 1307-1325. https://doi.org/10.1007/s11229-020-02793-y

Rigotti, M., Barak, O., Warden, M., Wang, X., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature, 497,* 585-590. https://doi.org/10.1038/nature12160

Roskies, A. (2021). Representational similarity analysis in neuroimaging: proxy vehicles and provisional representations. *Synthese, 199,* 5917-5935. https://doi.org/10.1007/s11229-021-03052-4

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind & Language, 22*(3), 246-269. https://doi.org/10.1111/j.1468-0017.2007.00308.x

Shea, N. (2013). Naturalising representational content. *Philosophy Compass*, 8(5), 496-509. https://doi.org/10.1111/phc3.12033

Shea, N. (2018). *Representation in cognitive science*. Oxford: Oxford University Press.

Shehzad, Z. & McCarthy, G. (2018). Category representations in the brain are both discretely localized and widely distributed. *Journal of Neurophysiology, 119*(6), 2256–2264. https://doi.org/10.1152/jn.00912.2017

Silberstein, M. & Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philosophy of Science, 80*(5), 958-970. https://doi.org/10.1086/674533

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*(1), 1-23. https://doi.org/10.1017/S0140525X00052432

van Gelder, T. (1992). Defining 'distributed representation'. *Connection Science, 4*(3-4), 175-191. https://doi.org/10.1080/09540099208946614

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience, 43*(1), 249-275. https://doi.org/10.1146/annurev-neuro-092619-094115

Weiskopf, D. (2021). Data mining the brain to decode the mind. In F. Calzavarini & M. Viola (Eds.), *Neural mechanisms: New challenges in the philosophy of neuroscience* (85-110). Cham: Springer International Publishing AG.

Willett, F. R., Deo, D. R., Avansino, D. T., Rezaii, P., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2020). Hand knob area of premotor cortex represents the whole body in a compositional way. *Cell, 181*(2), 396-409. https://doi.org/10.1016/j.cell.2020.02.043

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.